

INHERITED PREDISPOSITION TO IDIOPATHIC  
PULMONARY FIBROSIS IN THE NEWFOUNDLAND  
POPULATION

By

Ashar Pirzada

A thesis submitted to the School of Graduate Studies in partial fulfillment of  
the requirement for the degree of  
Master of Science

Discipline of Genetics  
Faculty of Medicine  
Memorial University of Newfoundland

May 2014

St. John's

Newfoundland

## **Abstract**

Idiopathic pulmonary fibrosis (IPF) is a late-onset disease characterized by inflammation and scarring of the lung parenchyma. 10-15% of IPF is attributed to genetic causes. The prevalence of familial pulmonary fibrosis (FPF) is up to 10x higher in Newfoundland & Labrador (NL) in comparison to other populations of European origin such as the United Kingdom (UK) and Finland. The five genes (*TERT*, *TERC*, *ABCA3*, *SFTPC* and *SFTPA2*) known to carry variants causing FPF have been screened in our NL cohort with no pathogenic variants found. This suggested there is/are novel variant(s) to be identified. Previous work done in this cohort utilized microsatellite genome-wide scans, fine-mapping/haplotyping and SNP genotyping to find loci associated with FPF. From these loci on chromosome 16 and 6, ten positional and functional candidate genes were previously sequenced with no pathogenic variants identified.

In this thesis, selection and sequencing of candidate genes from previously mapped loci is performed. Nine candidate genes were sequenced by Sanger sequencing in FPF Family R0942; however, no pathogenic variant was discovered out of 28 variants found. Also, genotyping was carried out on a common *MUC5B* rs35795950 promoter polymorphism that has been recently implicated with both sporadic and familial forms of pulmonary fibrosis (PF). A case-control analysis was carried out using 110 affected individuals and 277 healthy controls from the Newfoundland population. Results showed a significant association between rs35705950 genotypes and IPF. The odds ratio for individuals affected with IPF who were heterozygous and homozygous for the variant allele of this SNP were 5.4 (95% confidence interval, 3.3 to 9.6,  $P < .001$ ) and 12.2 (95%



confidence interval, 3.3 to 44.7,  $P < .001$ ), respectively. Furthermore, two of our FPF families (R0942 and R1136) showed familial segregation of the variant allele with the phenotype. In these families, all affected individuals were carriers of the variant T allele. Furthermore, a Simplified rapid Segregation Analysis (SISA) analysis demonstrated that the probability by chance that co-segregation of the variant T allele with PF in family R0942 was 1.56%.

This thesis supports the suggestion that the minor T allele of rs35705950 is likely a contributor to the pathogenesis of IPF in the NL cohort. The *MUC5B* gene encodes for a major gel-forming mucin macromolecule in respiratory secretions and is upregulated in other lung diseases. Further evidence of association is provided by tissue expression studies done through previous research.

## **Acknowledgements**

I would like to express my utmost gratitude to Dr. Michael Woods. Throughout the two years of my project, he provided a great amount of guidance, mentorship and patience for which I am very appreciative. I have learned much over these years and I could not have started or completed this project without his help. I would also like to thank my supervisory committee members, Dr. Bridget Fernandez and Dr. Terry-Lynn Young, for giving me additional insight and direction during the course of my project.

Thank you to all of the individuals who enrolled and participated in this study. Without your cooperation this project would not be possible. Thank you to the Atlantic Medical Genetics and Genomics Initiative and the Newfoundland and Labrador Lung Association for funding this project. Thank you to the Janeway Foundation and School of Graduate Studies for providing me with a stipend to carry out this research.

I would also like to thank Ms. Krista Mahoney, Ms. Erica Clarke and Mrs. Amanda Dohey for all of the help they have provided me. Whether it was answering any lingering questions I had or helping with my lab work, they were always willing to contribute to my project. Thank you to Mr. Fady Kamel for helping me make the transition into becoming a graduate student. Mr. Kamel helped me whenever I needed direction and his advice really provided me with a stable platform to begin my research. Finally, I would like to thank my family and friends for all of the support and encouragement they have given me throughout the course of my Masters project.

## **Table of Contents**

Abstract .....	ii
Acknowledgements .....	iv
List of Tables .....	viii
List of Figures .....	ix
List of Abbreviations .....	xi
1. Introduction .....	1
1.1 Interstitial lung disease .....	1
1.2 Idiopathic pulmonary fibrosis .....	3
1.3 The genetics of idiopathic pulmonary fibrosis.....	5
1.3.1 Familial pulmonary fibrosis.....	5
1.3.2 Documented FPF-causing genes.....	7
1.3.2.1 <i>TERT</i> and <i>TERC</i> .....	7
1.3.2.2 <i>SFTPC</i> and <i>SFTPA2</i> .....	10
1.3.2.3 <i>ABCA3</i> .....	12
1.3.3 Association of a <i>MUC5B</i> promoter polymorphism .....	13
1.4 The Newfoundland population .....	17
1.4.1 Population structure and genetic isolation.....	17
1.4.2 Familial Pulmonary Fibrosis in Newfoundland .....	19
1.5 Previous statistical work done .....	20
1.6 Previous work done.....	22
1.6.1 Participant Recruitment.....	22
1.6.2 Clinical Assessment .....	22
1.6.3 Genome-wide scan - microsatellite markers.....	25
1.6.4 Fine mapping – microsatellite markers.....	26
1.6.5 Genome-wide scan – SNP markers.....	27
1.6.6 Genes sequenced.....	29
1.6.7 Telomere assays .....	30
1.7 Hypotheses.....	31
1.8 Objectives .....	31
2. Materials and methods .....	33
2.1 Study Population .....	33
2.2 Candidate Gene Selection.....	34

2.3 Oligonucleotide Primer Design.....	35
2.4 Polymerase Chain Reaction / DNA Sequencing.....	36
2.5 Genomic Regions Analyzed / Characterizing Variants .....	40
2.6 <i>MUC5B</i> rs35705950 SNP genotyping.....	42
2.6.1 TaqMan genotyping procedure .....	42
2.6.2 Description of clinical variable parameters .....	44
2.6.3 <i>MUC5B</i> case-control statistics .....	45
3. Results.....	47
3.1 Clinical description and findings - Family R0942 .....	47
3.2 Candidate genes analyzed .....	49
3.2.1 Results from candidate gene sequencing - Chromosome 16 .....	51
3.2.1.1 <i>SOCS1</i> .....	51
3.2.1.2 <i>MSLN</i> .....	52
3.2.1.3 <i>TELO2</i> .....	54
3.2.1.4 <i>SNRNP25</i> .....	56
3.2.1.5 <i>IL17C</i> .....	57
3.2.2 Results from candidate gene sequencing - Chromosome 6 .....	64
3.2.2.1 <i>EDN1</i> .....	64
3.2.2.2 <i>CD83</i> .....	66
3.2.2.3 <i>DTNBP1</i> .....	67
3.2.2.4 <i>VEGFA</i> .....	69
3.3 <i>MUC5B</i> rs35705950 genotyping .....	72
3.3.1 Case vs. control analyses .....	72
3.3.2 Familial segregation of <i>MUC5B</i> promoter SNP rs35705950 .....	72
4. Discussion .....	78
4.1 Summary of the candidate gene analysis .....	78
4.2 The role of the <i>MUC5B</i> promoter SNP in IPF.....	80
4.3 Limitations of this study .....	83
4.4 Future work .....	85
5. Conclusion .....	87
References .....	88
Appendices.....	92
Appendix A: 2-point parametric linkage LOD scores on 500k SNPs for R0942, dominant model .....	92

Appendix B: Candidate gene tables for ROIs on chromosome 6p and 16p .....	93
Appendix C: Forward and Reverse Primer Sequences for All Genes Sequenced in Family R0942 .....	123
Appendix D: Thermocycler programs used .....	127
Appendix E: Setting up plate information for genotyping using SDS version 2.4 software.....	131
Appendix F: Ethanol precipitation of cycle sequencing reactions.....	132
Appendix G: All genes fully sequenced in the regions of interest on chromosome 6 and 16.....	133
Appendix H: Family R0942 pedigree with fine-mapping and haplotype analysis on chromosome 16p. Kamel, 2010 .....	134
Appendix I: Manuscript submitted to Thorax for publication on March 2013 .....	135
Appendix J: Publisher's permissions to use copyright materials.....	142

## **List of Tables**

Table 1.1: Criteria for diagnosis of IPF in absence of surgical lung biopsy. Adapted from ATS/ERS.....	23
Table 1.2: Microsatellite marker genome-wide scan results suggestive of linkage for A) Family R0942 and B) Family R0851. Adapted from Edwards, 2006 .....	26
Table 3.1: All genes fully sequenced in the regions of interest on chromosome 6 and 16 from this project .....	50
Table 3.2: Characteristics of all variants found by Sanger sequencing on chromosome ROIs .....	60
Table 3.3: Characteristics of all variants found by Sanger sequencing on chromosome ROIs .....	70
Table 3.4: Clinical variables and findings from rs35705950 <i>MUC5B</i> promoter SNP genotyping case vs. control study .....	75

## **List of Figures**

Figure 1.1: The Subtypes of Interstitial Lung Disease .....	2
Figure 1.2: Histopathologic features of IPF .....	4
Figure 1.3: Pedigree showing typical autosomal dominant disease inheritance.....	7
Figure 1.4: The various functions of telomerase.....	8
Figure 1.5: The crystal structure of the carbohydrate recognition domain and neck domain of rat surfactant protein A .....	12
Figure 1.6: MUC5B Expression in 33 Subjects with Idiopathic Pulmonary Fibrosis (IPF) and 47 Healthy Controls, Stratified According to rs35705950 Genotype and Smoking Status .....	16
Figure 1.7: Canadian provincial geography .....	18
Figure 1.8: 2-point parametric linkage LOD scores on 500k SNPs for R0942, dominant model .....	28
Figure 1.9: Average telomere lengths for lymphocyte and granulocyte cells in two affected individuals from family R0942 (partial pedigree). Kamel, 2010.....	32
Figure 2.1: R0942 with individuals genotyped for SNP-genome wide scan analysis.....	35
Figure 2.2: Family R0942 with DNA samples used for Sanger sequencing .....	38
Figure 2.3: SDS Allelic Discrimination genotype plot for <i>MUC5B</i> rs35705950 .....	43
Figure 3.1: Location of Trinity Bay on a map of Newfoundland .....	47
Figure 3.2: Family R0942 with identifying numbers.....	49
Figure 3.3: Physical location of genes sequenced within regions of interest on chromosomes 6 and 16.....	50
Figure 3.4: Segregation analysis of both variants found in <i>SOCS1</i> .....	52
Figure 3.5: Segregation analysis of three variants found in <i>MSLN</i> .....	54
Figure 3.6: Segregation analysis of three missense variants in <i>TELO2</i> .....	56

Figure 3.7: Segregation analysis of a 5'UTR variant and a 3'UTR variant in SNRNP25 and an exonic variant in IL17C .....	59
Figure 3.8: BioGPS human tissue gene expression profiles for SOCS1 / MSLN .....	62
Figure 3.9: A) BioGPS human tissue gene expression profile for TELO2. B) STRING TELO2 protein to protein interactions .....	63
Figure 3.10: Segregation analysis of an exonic variant found in EDN1 .....	65
Figure 3.11: Segregation analysis of three variants in exon 5 of CD83 .....	67
Figure 3.12: Segregation analysis of both variants found in DTNBP1 .....	68
Figure 3.13: BioGPS human tissue gene expression profiles for EDN1 / VEGFA.....	71
Figure 3.14: Excel bar graph comparing differences between genotype frequencies ....	73
Figure 3.15: Segregation analysis of <i>MUC5B</i> promoter SNP rs35705950 in two PPF families: R1136 and R0942.....	77



## **List of Abbreviations**

<i>ABCA3</i>	<i>ATP binding cassette protein A3</i>
AEC	Alveolar epithelial cell
ATS	American Thoracic Society
CGH	Comparative genomic hybridization
cM	Centimorgan
dbSNP	Global database of SNPs
dH <sub>2</sub> O	Distilled water
DLCO	Diffusing capacity of the lung for carbon monoxide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
DPLD	Diffuse parenchymal lung disease
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacetic acid
EtOH	Ethanol
FEV <sub>1</sub>	forced expired volume per second
FPF	Familial pulmonary fibrosis
FVC	Forced vital capacity
gDNA	Genomic DNA
HLOD	Heterogeneity log of odds score

HRCT	High resolution computed tomography
HWE	Hardy-Weinberg equilibrium
IIP	Idiopathic interstitial pneumonia
ILD	Interstitial lung disease
IPF	Idiopathic pulmonary fibrosis
LD	Linkage disequilibrium
LOD	Log of odds score
MM	Malignant mesothelioma
mRNA	messenger ribonucleic acid
<i>MUC5B</i>	<i>Mucin 5b</i>
NL	Newfoundland and Labrador
PCR	Polymerase chain reaction
PFT	Pulmonary function test
PMGP	Provincial Medical Genetics Program
RNA	Ribonucleic acid
ROI	Region of interest
SISA	Simplified rapid segregation analysis
<i>SFTPA2</i>	<i>Surfactant protein A2</i>
<i>SFTPC</i>	<i>Surfactant protein C</i>
SNP	Single nucleotide polymorphism
SNRNP	Small nuclear ribonucleoprotein
<i>TERC</i>	<i>Human telomerase RNA</i>
<i>TERT</i>	<i>Telomerase</i>

TLC	Total lung capacity
UIP	Usual idiopathic pneumonia
UK	United Kingdom
USA	United States of America
UTR	Untranslated region
WGS	Whole genome sequencing

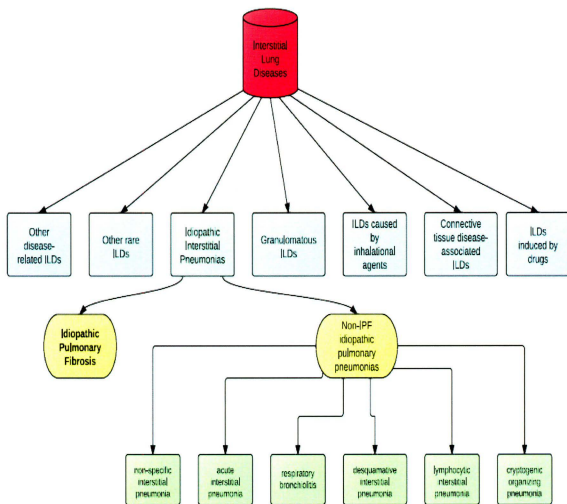
## **1. Introduction**

### **1.1 Interstitial lung disease**

Interstitial lung disease (ILD), also known as diffuse parenchymal lung disease (DPLD), describes a variety of lung diseases affecting the alveoli, airways and pulmonary interstitium. Estimates state that there are over 50 million people worldwide affected with one of the many ILDs (Maher, 2008).

ILDs can be broken down into seven major categories: 1) Idiopathic interstitial pneumonias; 2) Granulomatous ILDs (i.e. sarcoidosis); 3) ILDs caused by inhalational agents (i.e. asbestosis, silicosis); 4) Connective tissue disease-associated ILDs (i.e. scleroderma); 5) ILDs induced by drugs (i.e. bleomycin, amiodarone); 6) Other disease-related ILDs (i.e. some reno-pulmonary syndromes); 7) Other rare ILDs (i.e. lymphangioleiomyomatosis, alveolar proteinosis, langerhans' cell histiocytosis). Furthermore, the idiopathic interstitial pneumonias (IIP) category can be segmented into two groups: idiopathic pulmonary fibrosis (IPF) and non-IPF idiopathic pulmonary pneumonia. Non-IPF idiopathic pulmonary pneumonias are further divided into six groups: 1) non-specific interstitial pneumonia; 2) acute interstitial pneumonia; 3) respiratory bronchiolitis; 4) desquamative interstitial pneumonia; 5) lymphocytic interstitial pneumonia; 6) cryptogenic organizing pneumonia (Figure 1.1).

The different disease states can vary from being benign and running a very limited course to being progressively debilitating and lethal. The most common symptom of ILDs is shortness of breath- this is usually progressive (Peros-Golubicic and Sharma, 2006). Many of the disease states overlap in terms of clinical features, physiology



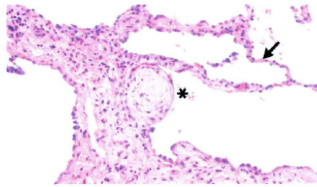
**Figure 1.1: The Subtypes of Interstitial Lung Disease**

measures, radiological and histological patterns (Alder et al., 2008). Therefore to make a diagnosis for one of the many ILDs, a multi-disciplinary approach involving clinicians, pathologists, and radiologists is required. This multi-disciplinary approach is now considered the “gold standard” for diagnosis of ILDs (Bradley et al., 2008). The specific ILD this thesis will concentrate on is IPF and the underlying genetic etiology.

## **1.2 Idiopathic pulmonary fibrosis**

Idiopathic pulmonary fibrosis is the most common type of all IIPs. IPF is a late-onset disease which is more frequent in males and generally occurs in the 5<sup>th</sup> to 7<sup>th</sup> decades of life (King et al., 2011). IPF has a debilitating prognosis - it presents as progressive dyspnea and fibrosis limited to the lung. Respiratory failure is common 3 - 5 years after diagnosis (ATS 2000). It is always associated with a histopathological appearance of usual interstitial pneumonia (UIP) when surgical lung biopsy is performed. The prevalence of IPF can vary depending on the population in question but typical estimates are between 1 - 24 cases per 100,000 (Raghu et al., 2006). Corticosteroids such as prednisone have been used as a therapy but there is little evidence of benefits to patients/clinical outcomes. A number of clinical trials using novel drugs (N-acetylcysteine, etanercept, bosentan) have resulted in some patients with mild-stage IPF showing positive change in pulmonary function tests (PFTs) after one year of treatment (King et al., 2011). However, all of these trials exclude patients with severe IPF so results are not generalizable to all IPF patients. Currently the only effective treatment that increases survival time for any individual with IPF is lung transplant.

To diagnose IPF, a systemic approach involving clinical evaluations, PFTs and high resolution computed tomography (HRCT) scans should be implemented (2000). If there is diagnostic confusion, lung biopsy is performed to provide a definitive diagnosis of IPF. The pathological changes are usually found in the subpleural parenchyma of the lung. As shown in Figure 1.2, honeycombing and fibroblastic foci are the hallmarks of typical UIP (Gross and Hunninghake, 2001). Diagnosis involves a multi-disciplinary team consisting of various clinicians, including respirologists, pathologists and radiologists.



**Figure 1.2: Histopathologic features of IPF. Reproduced with permission from and Hunninghake, 2001.** Copyright Massachusetts Medical Society. The asterisk represents a fibroblast focus; the arrow points to alveolar septa with little abnormality.

IPF has been associated with exposure to a variety of agents/factors including metal, wood, textile dusts, stone and sand, wood fire, cigarette smoking, certain livestock, agricultural offending agents, certain drugs (i.e. bleomycin) and other diseases (Peros-Golubicic and Sharma, 2006; King et al., 2011). Although these factors have been associated with IPF, there is a lack of consistency in these findings. Also, IPF cases in the absence of these factors are commonplace. Thus, the complete pathogenesis of IPF is still not completely understood. Previously, it was thought that fibrosis of the lung parenchyma was directly caused by chronic inflammation (Gross and Hunninghake, 2001). This school of thought stated that an initial stimulus led to chronic inflammation of the lung which in turn caused injury to the lung. Due to this, an aberrant inflammatory repair process ultimately caused fibrosis. However, this theory was not entirely correct for a number of reasons: anti-inflammatory drugs did not improve the outcome of IPF, inflammation is not always present histopathologically, and epithelial injury alone without inflammation can result in aberrant wound healing, extracellular matrix (ECM) remodeling and fibroblast foci formation (Selman et al., 2001).

The current prevalent theory for pathogenesis of IPF involves the combination of genetic/environmental factors and alveolar epithelial cell (AEC) injury as a primary step. Briefly, the susceptible lung, by virtue of genetic factors, shortened telomeres or reduced ability to regenerate, is affected by repetitive injuries (smoking, offending agents etc). This ultimately results in the death of many alveolar type I and II epithelial cells (King et al., 2011). In turn, wound clots and other AECs are aberrantly activated, producing a number of growth factors and chemokines which migrate to the site(s) of lung injury. All of these factors are involved in the differentiation of fibroblasts to myofibroblasts. This is indicative of the breakdown of normal lung repair processes (King et al., 2011). Myofibroblasts, which are differentiated cells with potent profibrotic ability, leads to increased and irreversible fibrosis by depositing ECM as well as further AEC death (Waghray et al., 2005). These theories and observations illustrate that multiple molecular mechanisms may be implicated in pathogenesis in that the inflammatory pathway and/or the aberrant epithelial pathway can lead to PF (King et al., 2011).

From these previous theories and observations, IPF is believed to be a complex disease with more than one environmental and/or genetic factor involving a variety of physiological pathways to produce the phenotype, the primary focus of this thesis will be determining the genetic susceptibility of IPF in the NL population.

### **1.3 The genetics of idiopathic pulmonary fibrosis**

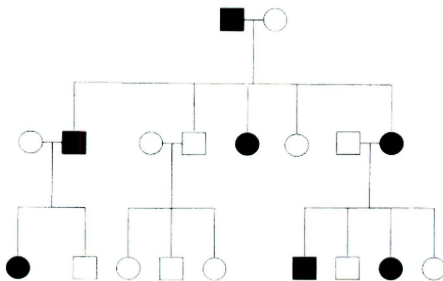
#### **1.3.1 Familial pulmonary fibrosis**

Familial pulmonary fibrosis (FPF) is defined as the occurrence of IPF within two or more members of a family- these individuals must be 1<sup>st</sup> or 2<sup>nd</sup> degree relatives with a



common ancestor. The clinical presentation of FPF (symptoms, medical examination, PFT levels, radiological and pathological findings) generally matches that of IPF- the only difference is that FPF has an earlier age of onset than IPF by approximately 8-12 years (Tsakiri et al., 2007). Although FPF is primarily a late-onset disease, there have been recent rare findings of FPF in three siblings diagnosed very early in adulthood as well as rapid progressive ILD in young children not more than 2 years of age (Vece et al., 2012).

Anywhere between 2-19% of patients with IPF report having a 1<sup>st</sup> degree relative with similar symptoms (van Moorsel et al., 2010). In affected families, FPF typically displays an autosomal dominant inheritance pattern (Figure 1.3) with reduced penetrance. An autosomal dominant inheritance pattern implies that if an affected mutation carrier has offspring, an average of 50% of the offspring of the mutation-carrier will carry the disease-causing variant. Penetrance refers to the correlation between disease genotype and phenotype. If a disease is fully penetrant, every individual with the disease-causing genetic variant will present with the disease phenotype. In a disease that has reduced penetrance, such as FPF, there are individuals who may carry the disease-causing genetic variant but will not present with the disease phenotype during their lifetime. Furthermore, phenocopies can also be present in families with FPF. Phenocopies are defined as the same phenotype being displayed by different individuals due to differing genetic and/or environmental causes. These genetic factors complicate the ability to identify a haplotype and pathogenic variants that segregate with the FPF phenotype (Strachan and Read, 1999).



**Figure 1.3: Pedigree showing a typical autosomal dominant inheritance pattern**

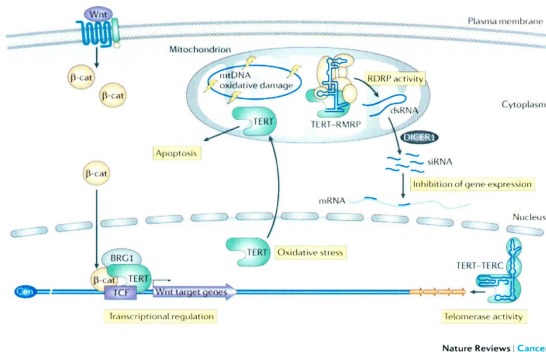
### **1.3.2 Documented FPF-causing genes**

To date, causal variants in the genes *TERT*, *TERC*, *SFTPC*, *SFTPA2* and *ABCA3* account for 2-20% of FPF depending on the population studied (Armanios et al., 2007; Nogee et al., 2001; Wang et al., 2009). Out of these known FPF-causing genes, genetic variants in *TERT* and *TERC* are the most common.

#### **1.3.2.1 *TERT* and *TERC***

*TERT* is a gene located on chromosome 5p15.33 at position 1,253,287 - 1,295,162 (hg19). *TERT* codes for the telomerase enzyme. Telomerase has both a reverse transcriptase catalytic component (also known as *hTERT*) and a functional RNA component (*TERC*) (Diaz de Leon et al., 2010). As shown in Figure 1.4, telomerase is involved in many pathways and thus has a variety of important physiological functions in

the body. Telomerase functions as a transcriptional regulator of the Wnt- $\beta$ -catenin pathway, addition of telomeric repeats to stabilize chromosomes, a regulator for apoptosis in the mitochondria, production of small interfering RNAs that can inhibit gene expression, and a determinant of oxidative stress in the cell (Martinez and Blasco, 2011).



**Figure 1.4: The various functions of telomerase. Reprinted with permission from Martinez and Blasco, 2011 (Appendix J).**

The main function of telomerase is the addition of telomeric repeats to the end of chromosomes. This stabilizes and slows down the progressive shortening of chromosomes that occurs with each cell division (Armanios et al., 2007). Once the telomeres reach a critical size, the cell stops its normal cell cycle and/or apoptosis (programmed cell death) is initiated. Thus, the length of telomeres is a limiting factor for the capacity of tissues to replicate. If chromosomes have short telomeres, each cell division will lead to the telomere reaching the critical size in a shorter time period and

thus apoptosis/ inhibition of the cell cycle will occur earlier than average. This is one of the main reasons abnormal telomere length is associated with age-related disease and disease involving tissues with a high rate of turnover, such as the lung (Armanios et al., 2007).

Due to its various functional roles in the body, defective telomerase has been shown to result in a variety of disorders such as dyskeratosis congenita, aplastic anemia, coronary artery disease, liver cirrhosis and/or PF. One of the disorders, dyskeratosis congenita, is a multi-system disorder that can have a variety of clinical features such as abnormal skin pigment, bone marrow failure, lung and liver fibrosis, premature graying of the hair, osteoporosis and mental retardation, among other varying features (Bessler et al., 2010).

FPF is more common in the 5<sup>th</sup> decade of life and beyond. As well, the epithelial tissue in the lung displays a relatively high turnover rate in comparison to other tissues. We also know that abnormal telomere length is associated with later onset disease and disease involving tissues with a high rate of turnover, and a number of studies have found genetic variations in *TERT* and *TERC* to be the most common molecular cause of FPF (Armanios et al., 2007; Tsakiri et al., 2007; Alder et al., 2008; Diaz de Leon et al., 2010). Considering the various functions that telomerase has in the cell, it is surprising that heterozygous mutations in *TERT* and *TERC* commonly cause a specific pulmonary fibrosis phenotype instead of a multi-systemic phenotype.

In a study using the Vanderbilt Familial Pulmonary Fibrosis Registry for recruitment, 6/73 FPF probands has heterozygous mutations in *TERT* or *TERC* (Armanios et al., 2007). These mutations also segregated with FPF in all six families and were not

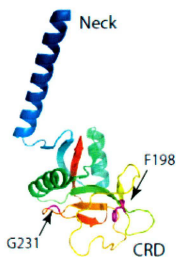
present in 623 unaffected subjects. Furthermore, telomere length in lymphocytes was significantly lower in probands and asymptomatic mutation-carriers than relatives without the mutation (Armanios et al., 2007). From this study, it was evident that mutant telomerase is associated with FPF. Further studies have also shown that heterozygous mutations in *TERT* or *TERC* cause FPF and over time confer a large increase in susceptibility to adult-onset FPF with the phenotype presenting earlier in successive generations (Tsakiri et al., 2007).

#### 1.3.2.2 *SFTPC* and *SFTPA2*

Two surfactant protein genes, *SFTPC* and *SFTPA2*, have also been associated with IPF. Surfactant is a pulmonary-specific viscous fluid consisting of various proteins and phospholipids. Surfactant is produced by Type II alveolar epithelial cells and its main function is to reduce surface tension by lining the epithelial air space. This is essential in allowing lungs to expand and contract with ease by stabilizing alveoli as well as to prevent the collapse of the lung due to abnormal pressure changes. Surfactant proteins such as *SFTPA2* and *SFTPC* are integrally involved in these processes. These surfactant protein components are also part of a homologous family of immune-defense proteins called collectins. Mouse models have shown that without subunits of surfactant such as *SFTPA/A2*, mice are susceptible to pulmonary infections and disease (Wang et al., 2009). Due to lung surfactant being made up of a variety of phospholipids, surfactant proteins and cholesterol, any harm/change in either one these components can have disruptive effects in normal lung function. Certain surfactant proteins were first documented in association with IPF in 1991 when a study showed that there were elevated levels of surfactant protein A in IPF patient bronchoalveolar lavage in comparison to healthy

controls (McCormack et al., 1991). Other studies have focused on using surfactant proteins as biomarkers and/or prognostic factors in IPF with variable results (Takahashi et al., 2000; Greene et al., 2002).

Subsequently, there have been recent studies that have found *SFTPC* mutations as a cause for both familial and sporadic forms of IPF. One study found genetic variations in *SFTPC* in 11/34 patients with ILD (Nogee et al., 2001). The main variant found was a splicing mutation (c.460+1G>A) that resulted in skipping of exon 4 and reduced expression of *SFTPC*. Also, this variation was absent from 100 control chromosomes (Nogee et al., 2001). Another study found an *SFTPC* exon 5 mutation in two FPF families and seven sporadic IPF cases which results in a missense amino acid change that can affect peptide processing (Chibbar et al., 2004). Furthermore, variants in *SFTPA2* have also been associated with IPF. Wang et al. found two rare variants in *SFTPA2* (c.692G>T, p.G231V; c.593T>C, p.F198S) which caused FPF in two separate families (2009). Both mutations disrupt tertiary structure of a carbohydrate recognition domain in the protein (Figure 1.5). Site-directed mutagenesis and human lung epithelial cell lines were used to see if the mutations effected synthesis and secretion of the protein. In comparison to wild-type protein, the *SFTPA2* p.G231V and p.F198S variants were expressed in cell lines very poorly and tested negative for SFTPA2 protein in cell lines (Wang et al., 2009). Thus, this study demonstrated that these *SFTPA2* germline variants disrupt protein trafficking/function and result in FPF.



**Figure 1.5: The crystal structure of the carbohydrate recognition domain, neck domain of rat surfactant protein A, and positions corresponding to codons 198 and 231 of the human sequence. Reprinted with permission from Wang et al., 2009 (Appendix J).**

Since there have been separate genetic variations identified in *SFTPC* and *SFTPA2* genes that account for a small proportion of FPF in various populations around the world, these surfactant protein genes have become well-documented FPF-causing genes (Nogee et al., 2001; Wang et al., 2009; Ono et al., 2011).

#### 1.3.2.3 *ABCA3*

The *ABCA3* gene, ATP-Binding Cassette sub-family A member 3, is another gene that can harbor FPF-causing variants as demonstrated by previous research (Young et al., 2008). The protein encoded by the *ABCA3* gene is a membrane protein part of the ATP-binding cassette transporter family. ATP-binding transporter proteins function by utilizing the energy of ATP to actively transport various substrates across a membrane. *ABCA3* is expressed exclusively in lung tissue and functions as part of a transmembrane transporter of the phospholipid components of surfactant (Yamano et al., 2001). Since *ABCA3* is

needed to maintain pulmonary surfactant phospholipid balance, genetic variants modifying the activity of this gene can have serious consequence in lung-related disorders (Young et al., 2008). Interestingly, the processing and secretion of surfactant protein C is dependent on *ABCA3*, highlighting its involvement in a biological pathway known to be involved in some cases of IPF.

Genetic variants in *ABCA3* have been shown to cause IPF in adolescents as well as increase the severity of FPF in patients with *SFTPC* variants (Young et al., 2008; van Moorsel et al., 2010). The first documented case of *ABCA3* variants causing PF was reported in 2008. A 15 year-old Caucasian boy of European descent presented with symptoms of lung disease and was diagnosed with IPF based on usual interstitial pneumonia histology (Young et al., 2008). Genetic analysis did not find variations in any known FPF genes but novel heterozygous *ABCA3* variants were discovered (c.-28A>G; c.3765C>G). Furthermore, there have been other *ABCA3* variants found in unrelated children with mild lung disease, implying that *ABCA3* could also be predictive of other lung-disease phenotypes (Bullard et al., 2005, Crossno et al., 2010). Due to *ABCA3*'s role in processing surfactant protein C, one study tested the hypothesis that potential *ABCA3* variants could modify the severity of IPF in patients with *SFTPC* variants (van Moorsel., 2010). This study showed that sisters carrying an *ABCA3* variant (c.3784A>G; p.S1262G) presented with FPF 15 years earlier than affected patients without this variant but the effects of this variant on FPF are still not yet fully understood (van Moorsel., 2010).

### **1.3.3 Association of a *MUC5B* promoter polymorphism**

Recently, a variant (rs35705950: dbSNP; <http://www.ncbi.nlm.nih.gov/projects/SNP/>) located in a promoter region 3kb upstream of a major gel-forming mucin

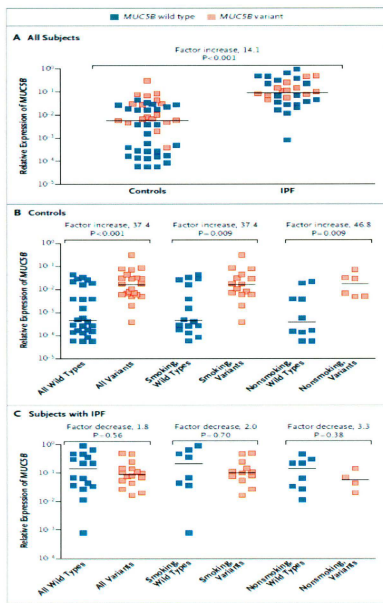


gene, Mucin 5b (*MUC5B*), that is associated with IPF in sporadic and familial forms in populations in the USA was identified (Seibold et al., 2011). Specifically, the minor T allele was present in 38% of IPF subjects and 34% of Familial Interstitial Pneumonia (FIP) subjects in comparison to 9% of controls (Seibold et al., 2011). All documented FPF-causing genetic variants are in coding sequences that result in a measurable pathogenic change in the protein. The rs35705950 variant is an intergenic variant with no direct effect on protein and thus is harder to prove as causative.

However, it was demonstrated that unaffected individuals with at least one copy of the minor T allele expressed *MUC5B* protein in lung tissue 37.4 times more than unaffected individuals with the wild-type G allele (Seibold et al., 2011). This analysis was done using biopsied lung tissue samples from unaffected individuals with at least one copy of the T allele and unaffected individuals with the wild-type G allele. The results suggest that gene expression is primarily altered by having one copy of the variant T allele and not due to presence of the phenotype. *MUC5B* expression in the lung for affected individuals was also 14.1 times higher than in unaffected individuals (Seibold et al., 2011). To see if smoking modified the effect of *MUC5B* expression, lung tissue from smokers and non-smokers was also examined to test for differences in *MUC5B* protein expression. Smoking did not modify the expression of *MUC5B* in the lung in healthy controls or affected individuals (Figure 1.6). Furthermore, the variant is in a known 5' promoter region of *MUC5B* and is conserved across primate species. Aberrant *MUC5B* expression has profound effects in the lung and has been previously shown to be involved in other lung disease, being upregulated in chronic pulmonary disease progression (Kirkham et al., 2008) and downregulated in cystic fibrosis airway secretions (Henke et

al., 2004). The association of this single nucleotide polymorphism (SNP) with IPF has since been confirmed by one other group in the same issue of the original paper (Zhang et al., 2011) and hence seems to play a role in the development of IPF. This *MUC5B* promoter SNP has also recently been associated with IPF-specific pathways rather than diseases that can exhibit pulmonary fibrosis such as systemic sclerosis and sarcoidosis (Stock et al., 2013). These findings suggest that there is a distinct genetic susceptibility that specifically contributes to IPF and is less likely to involve the immunological and/or inflammatory mechanism than pathogenesis via a mucus-related mechanism (Stock et al., 2013). Furthermore, the *MUC5B* promoter variant was associated with a less severe decline in patients' pulmonary function tests and slower disease progression with compared to patients with systemic sclerosis and sarcoidosis. Another study conducted using a Caucasian American cohort showed similar findings in that the *MUC5B* promoter variant was associated with significantly improved survival in IPF (Peljto et al., 2013).

Both of these recent findings suggest IPF and the *MUC5B* promoter SNP may be linked with a distinct disease pathogenesis. It is possible that the IPF phenotype is heterogeneous and that the *MUC5B* promoter polymorphism may be responsible for this (Peljto et al., 2013).



**Figure 1.6: MUC5B Expression in 33 Subjects with IPF and 47 Healthy Controls, Stratified According to MUC5B Promoter Single-Nucleotide Polymorphism (SNP) (rs35705950) Genotype and Smoking Status. Reproduced with permission from Seibold et al (2011). Copyright Massachusetts Medical Society.**

**Panel A:** the distribution of MUC5B expression among subjects with wild-type or heterozygous rs35705950 genotypes of MUC5B according to the presence or absence of IPF. **Panel B:** MUC5B expression among all controls, controls who smoked, and controls who did not smoke, according to rs35705950 genotype. **Panel C:** comparison of MUC5B expression in all subjects with IPF, those who smoked, and those who did not smoke, according to rs35705950 genotype. In all the panels, the horizontal lines indicate group medians, and the expression of MUC5B is shown in relation to the expression of the glyceraldehyde 3-phosphate dehydrogenase gene (GAPDH).

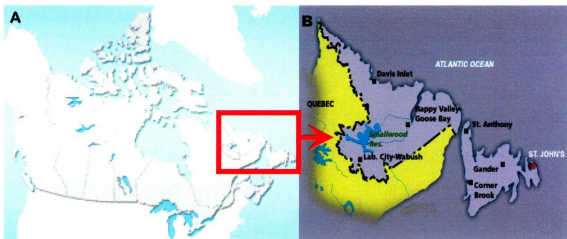
## **1.4 The Newfoundland population**

### **1.4.1 Population structure and genetic isolation**

The province of Newfoundland and Labrador (NL) consists of both an island (Newfoundland) as well as a continental landmass (Labrador) and is located on the east coast of Canada (Figure 1.7). According to Statistics Canada 2012, NL has a population of approximately 515,000. Since immigration started in the 1600s and was at its highest point in the mid-1700s, Newfoundland is a very young population (only ~20 generations old). This population is unique since it remains spread out in geographically isolated communities- approximately 60% of the current population live in communities with a total population of 2500 or less (Rahman et al., 2003) and there is only one urban centre with a population greater than 50,000. These communities, almost all of which were originally settled by the English and Irish, are long-standing, established settlements with very little or no migration at this time (Bear et al., 1987). Furthermore, most marriages occurred between individuals with similar religious practices and there was little to no admixture between individuals of differing religions. The chief means for the increase in the population of NL after 1800 is natural increase and 98% of the NL population in 2003 has English or Irish ancestry, illustrating its homogenous nature (Rahman et al., 2003).

Due to the geographical and religious isolation, minimal inwards migration and small population size that arose from a limited founder population, Newfoundland is considered a series of small genetic isolates. In these genetic isolates with a small number of founders, there is less genetic variation and some rare disease alleles may be present at a higher frequency, resulting in an enrichment of certain diseases. It can be easier to detect certain rare disease alleles in isolated populations since fewer haplotypes are

segregating through an isolated population. This is an overrepresented amount of Mendelian disorders in Newfoundland compared to other Canadian provinces after accounting for population size, highlighting its importance for studying genetic disease (Rahman et al., 2003).



**Figure 1.7: Canadian provincial geography.** A) Area shaded in grey highlights Canada. White areas to the bottom and left highlights USA. White areas to the top highlights Greenland. B) The province of Newfoundland and Labrador. Newfoundland is the island portion with St. John's as the capital city; Labrador is the continental region bordering Quebec (yellow).

In the Newfoundland population, a founder effect has been documented in a number of disorders such as Lynch Syndrome (Stuckless et al., 2007), hemophilia A (Xie et al., 2002), Bardet-Biedel Syndrome (Young et al., 1999) and Arrhythmogenic right ventricular cardiomyopathy type 5 (Merner et al., 2008). A founder effect is seen when a limited genetic diversity forms in a population due to a small group of individuals (i.e. founders) that become isolated for a number of reasons. Due to isolation, the gene pool of these “founders” is disproportionately more common in the resulting population and

genetic diversity is reduced. For instance in Newfoundland, many Europeans (primarily of English and Irish descent) who migrated into an outpost stayed in the same geographical region with little to no migration, creating clusters of small founder populations distributed across the entire island. Since FPF is a rare disease with a strong genetic component that has a high prevalence in the Newfoundland population compared to others, the possibility of a founder mutation cannot be ruled out as being a cause. Alternatively, there may be heterogeneous genetic causes that contribute to the high prevalence of FPF in Newfoundland.

#### **1.4.2 Familial Pulmonary Fibrosis in Newfoundland**

The prevalence of FPF in the Newfoundland population is over 12 cases per 100,000. This is considerably high when compared to most population groups around the world. For instance, populations in the USA and the United Kingdom (UK) had genetic transmission of PF in only 1-3 cases per 100 000 (Gribbin et al., 2006; Raghu et al., 2006). In Finland, a country with a well-known founder population which also contains many rare disease phenotypes, the prevalence of FPF is less than one per 100 000 (du Bois, 2006). Since the Newfoundland population is enriched for FPF, it is an excellent population to study to determine the genetic variations that cause this disease.

To date, over 80 patients from more than 30 Newfoundland families have been diagnosed with FPF (Woods, unpublished data). The number of affecteds per family ranges from only two affected individuals in a family to a large, six-generation family with ten affected individuals. The proband in each of these families as well as all sporadic IPF probands were previously screened for genetic variants in the following genes: *TERT*, *TERC*, *SFTPC*, *SFTPA2*, and *ABCA3*. Variants in *TERT* were identified in eight affected

individuals. Seven of these individuals were from two unrelated families with FPF and one individual with sporadic IPF. These variations (c.2594G>A; p.Arg865 His and c.2648T>G; p.Phe883Cys) were IPF-causing based on previous linkage findings by another research group (Tsakiri et al., 2007) and bioinformatics analysis. No other disease-causing variants were found in the remainder of FPF cases.

### **1.5 Previous statistical work done**

The 1000 Genomes Project shows that each individual carries ~250 to 300 loss-of-function variants in annotated genes (Altshuler et al., 2010). Thus, it is extremely important to be able to correctly exclude non-causative genetic variations and narrow down the regions of interest in order to find a disease-causing variant. Linkage analysis is used in order to identify genetic regions that are linked with a disease in a particular family or families when a mode of inheritance is known (Greenberg et al., 1998).

Genetic linkage analysis is based on the assumption that genes physically close together on a chromosome tend to be inherited together, or linked, when passed on through generations from parents to offspring. A genome-wide scan using SNPs or microsatellite markers is carried out in affected and unaffected family members. Since recombination occurs in germ cells that are passed on to offspring, the parental origin of different alleles is easily traced. When genetic markers are closer together, they are more likely to be inherited together than markers much further apart. This probability for a recombination between two genes is known as the recombination fraction, theta ( $\theta$ ).

Linkage analysis measures the expected versus observed recombination events given the parameters for genetic inheritance modes, penetrance of disease and frequency of genetic variations to determine region(s) linked with disease (Lathrop et al., 1984). Generally, genes with a pathogenic variant have a marker or multiple markers close by that are in linkage disequilibrium (LD) with this variant. The pathogenic variant that is in LD with one or more markers segregates in affected individuals. In linkage analysis, all regions, including regions of LD in a single family or multi-family analysis are shown by a specific log of odds (LOD) score for each marker.

LOD scores indicate the likelihood of whether loci are linked or not. LOD scores are quantified based on logarithms (10 times) and can be positive or negative values. There is one assumption used for the calculation of LOD scores: two genes are tightly linked with no recombination events occurring between them, indicated by  $\theta = 0$ . LOD scores can also be calculated for different theta values. For instance, a LOD score of 3.0 ( $\theta = 0$ ) indicates that the marker(s) in question has a 1000:1 likelihood of being linked with a pathogenic variant or another marker- this is considered significantly linked. A LOD score of -1.0 ( $\theta = 0$ ) indicates that the marker(s) in question has a 100:1 likelihood of not being in LD with a pathogenic variant or another marker. LOD scores are used in both two-point and multipoint linkage analyses. In two-point linkage analyses, there is usually one marker plus the disease being compared in order to calculate the LOD score; whereas, multipoint linkage analysis compares a number of markers to calculate the LOD score. Horne et al (2003) suggested multipoint linkage analysis is less sensitive to incorrect allele frequencies but can be prone to statistical loss of power. Nevertheless,



reasonable concordance is found between the different analysis methods (Horne et al., 2003). SNP genome-wide scans and both two-point and multi-point linkage analyses were previously carried out on informative Newfoundland FPF families including Family R0942.

## **1.6 Previous work done**

### **1.6.1 Participant Recruitment**

Affected individuals were diagnosed using the American Thoracic Society diagnostic criteria (2000) for probable IPF (Table 1.1). These individuals were asked to participate in this study by the respirologist initially examining the patients. Interested individuals were contacted by the research nurse coordinator for the study, Mrs. Barbara Noble, and written consent was obtained for the use of family/medical histories and biological samples for research (HIC #2.26). After the proband was enrolled in the study, all 1<sup>st</sup> and 2<sup>nd</sup> degree relatives were contacted by Mrs. Noble and asked to participate in the study also. Dr. Bridget Fernandez, the principal investigator, and Mrs. Barbara Noble, then used this information to construct family pedigrees. For all study participants, venous blood was drawn from affected and unaffected participants and whole genomic DNA was extracted by Mrs. Amanda Dohey or Ms. Krista Mahoney (Memorial University).

### **1.6.2 Clinical Assessment**

A methodical procedure was used to diagnose all affected individuals enrolled in this study. Definitive diagnosis of IPF is only possible via surgical lung biopsy. However, this is an invasive procedure with a mortality rate of 4.3% after 30 days (Park et al., 2007).

Thus, lung biopsy to definitively confirm diagnosis was only available in a minority of patients. Due to the invasive nature of lung biopsy, the ATS/ERS criteria for diagnosis of IPF in absence of surgical lung biopsy was used for all diagnoses (Table 1.1).

**Table 1.1: Criteria for diagnosis of IPF in absence of surgical lung biopsy. Adapted from ATS/ERS.**

### **Major Criteria**

Exclusion of other known causes of ILD such as certain drug toxicities, environmental exposures, and connective tissue diseases  
Abnormal pulmonary function studies that include evidence of restriction (reduced VC, often with an increased FEV<sub>1</sub>/FVC ratio) and impaired gas exchange [increased P(A-a)O<sub>2</sub>, decreased PaO<sub>2</sub> with rest or exercise or decreased DL<sub>CO</sub>]

Bibasilar reticular abnormalities with minimal ground glass opacities on HRCT scans  
Transbronchial lung biopsy or BAL showing no features to support an alternative diagnosis

### **Minor Criteria**

Age > 50 yr  
Insidious onset of otherwise unexplained dyspnea on exertion  
Duration of illness > 3 mo  
Bibasilar, inspiratory crackles (dry or “Velcro”-type in quality)

---

*Definition of abbreviations:* BAL = bronchoalveolar lavage; DL<sub>CO</sub> = diffusing capacity of the lung for CO; HRCT = high-resolution computerized tomography; ILD = interstitial lung disease; P(A-a)O<sub>2</sub> = alveolar-arterial pressure difference for O<sub>2</sub>; VC = vital capacity.

\* Reprinted by permission from Reference 60.

<sup>1</sup> In the immunocompetent adult, the presence of all of the major diagnostic criteria as well as at least three of the four minor criteria increases the likelihood of a correct clinical diagnosis of IPF.

Each participant in the study completed a medical history form, a family history questionnaire and an occupational exposure questionnaire. All participants were offered

clinical assessment which initially included a physical examination, PFTs and HRCT of the chest. PFTs include testing for forced vital capacity (FVC), forced expired volume/sec (FEV<sub>1</sub>) and diffusing capacity of the lung for carbon monoxide (DLCO). These PFTs were examined by Dr. George Fox, a respirologist (Memorial University).

HRCTs of the chest were also performed on patients who consented and the results were interpreted blindly by two radiologists, Dr. Rick Batia and Dr. Eric Sala (Memorial University). Scoring was based on the Royal Brompton Hospital scoring system (Wells et al., 2003). Concisely, each radiologist scored the images at five levels. Each level was given a score based on the proportions of ground-glass attenuation, reticular abnormalities and the coarseness of these reticular abnormalities (Wells et al., 2003). This scoring method allowed the radiologists to label each case as definite IPF, probable IPF, possible IPF or unaffected.

In the event that a lung biopsy was performed, a diagnosis of UIP is the only indicator that a patient that was definitely affected- this was performed via a single blind procedure with respect to the pathologist. Overall, with the combination of family pedigrees, medical history, PFTs, HRCTs and tissue biopsies, a panel of physicians/ researchers reviewed the information to make a final diagnosis and classify each individual as definitely affected, probably affected, possibly affected or unaffected.

The clinical classifications differed from the classifications used in this thesis in a number of ways. For an individual to be classified definitely affected, the individual must present with UIP on lung biopsy. For an individual to be classified probably affected, the individual must satisfy the ATS/ERS criteria (Table 1.1). Nonetheless, for the purpose of

statistical analysis in this study, all individuals that were clinically classified as definitely or probably affected were assigned a classification of definitely affected.

### **1.6.3 Genome-wide scan - microsatellite markers**

A genome-wide scan was previously done (Ms. Laura Edwards, Memorial University) on two families, R0851 and R0942, using a 10cM resolution human mapping set consisting of 382 microsatellite markers (Edwards, 2006). Microsatellite markers for the genome-wide scan covered the 22 autosomal chromosomes and excluded the X chromosome since the majority of FPF exhibits an autosomal dominant inheritance pattern (Steele and Brown, 2007).

In family R0942, 87% of the microsatellite markers were successfully genotyped. Using this data, two-point linkage analysis was done. Two-paired linkage analysis resulted in none of the markers with a LOD score of  $\geq 3.00$ , indicating that no markers showed statistically significant linkage to a disease locus. There were 11 markers that had a LOD score ranging from  $\geq 1.00$  to  $\leq 3.00$ , indicating that 11 markers showed suggestive linkage to a disease locus (Table 1.2a). The experimental maximum LOD score for family R0942 was 2.19 for marker D16S423, located at 16p13.3. Since simulation analysis revealed that the theoretical maximum LOD score for these families is 2.10 at  $\theta=0$  (due to the size and structure of the family), which is slightly lower than the LOD score obtained for marker D16S423, fine-mapping was done (section 1.6.4) to further analyze this region on chromosome 16 (Kamel, 2010).

In family R0851, 84% of the microsatellite markers were successfully genotyped. Using this data, two-point linkage analysis was done. None of the markers had a LOD score of  $\geq 3.00$ , indicating that no markers showed statistically significant linkage to a

disease locus. There were 4 markers that had a LOD score ranging from  $\geq 1.00$  to  $\leq 3.00$ , indicating that 4 markers showed suggestive linkage to a disease locus (Table 1.2b). This region on chromosome 16p 13 was analyzed further after a SNP genome-wide scan was performed using 5 families.

**Table 1.2: Microsatellite marker genome-wide scan results suggestive of linkage for A) Family R0942 and B) Family R0851. Adapted from Edwards, 2006.**

<b>A</b>	<b>Marker</b>	<b>LOD Score</b>	<b>B</b>	<b>Marker</b>	<b>LOD Score</b>
	D2S347	1.83*		D1S2890	1.09
	D2S338	1.33		D1S207	1.26
	D4S412	1.11		D4S405	1.20
	D6S309	1.11		D4S1592	1.32
	D6S470	1.41			
	D6S289	1.11			
	D7S531	1.11			
	D7S517	1.11			
	D9S287	1.17			
	D16S423	2.19*			
	D20S195	1.17			
	<b>TOTAL: 11</b>	<b>-</b>		<b>TOTAL: 4</b>	<b>-</b>

#### **1.6.4 Fine mapping – microsatellite markers**

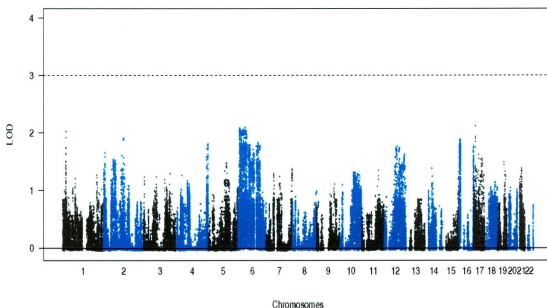
Using the results from the initial 10cM microsatellite marker genome-wide scan on family R0942, fine-mapping was carried out by Ms. Laura Edwards and Mr. Fady Kamel (Memorial University) using additional microsatellite markers between D16S3084 and D16S3046 (Kamel, 2010). This was done using up to 10 microsatellite markers on chromosome 16p for each of these six families: R0851, R0892, R0896, R0942, R1136

and R1487. After the successful genotyping of microsatellite markers, haplotypes were determined assuming least number of recombinations for each family. Focusing on family R0942, there was a haplotype that was determined to be segregating with FPF. This observed haplotype is on chromosome 16p- specifically, from the telomeric end (top) of 16p to marker D16S3103. This is a 17.38Mb region that contains 386 genes in total, including known and hypothetical genes (Kamel, 2010).

### **1.6.5 Genome-wide scan – SNP markers**

In addition to the microsatellite marker genome-wide scan performed by Ms. Laura Edwards on families R0942 and R0851, a SNP marker genome-wide scan was performed using five families: R0851, R0892, R0896, R0942 and R1136. The SNP marker genome-wide scan was done using a 610 Quad Illumina array (Illumina, Inc., California, USA) which genotypes over 550,000 SNP markers distributed across the genome. Affected and unaffected individuals from each family were genotyped using the 610 Quad Illumina array. This SNP marker genome-wide scan was done by the Genome Quebec Innovation Centre (Quebec, Canada).

To determine if any of the genotyped SNP marker(s) were linked to a causal variant, two-point and multipoint linkage analysis was performed (Kamel, 2010). All of these statistical analyses and tests were performed by Dr. Marie Pierre-Dubé (Université de Montréal, Montreal Heart Institute, [www.statgen.org/](http://www.statgen.org/)) and her group. Additionally, the results of the statistical analyses were analyzed by Dr. Michael Woods, Mr. Fady Kamel and me (Manhattan plot shown in Figure 1.8).



**Figure 1.8: 2-point parametric linkage LOD scores on 500k SNPs for R0942, dominant model.** This is a Manhattan plot received by Mr. Fady Kamel, 2010. This data and the accompanying SNP genome-wide analysis for R0942 were re-visited by me in order to confirm and fine-map the ROIs.

Using the data from the five families, there were only two markers that produced HLOD (heterogeneity LOD) scores over 3.00 using two-point parametric linkage analysis. HLOD scores are calculated by a formula that combines LOD scores for all of the families used in the analysis. This method is useful if there is a suspected rare variant or founder variant accounting for the phenotype in all families. The two markers that produced HLOD scores over 3.00 using two-point parametric linkage analysis were marker rs942631 (HLOD = 3.22) and marker rs3130922 (HLOD = 3.15), located on chromosomal regions 6p24.3 and 6p21.3 respectively (Kamel, 2010). Due to these HLOD scores being above 3.00, indicating statistically significant linkage to a disease locus,

these chromosomal regions were analyzed further and two genes from this region, *snoRNA66* and *PIN1L*, were selected for full gene sequencing (section 1.6.6).

Multipoint parametric linkage analysis only produced one region of interest (ROI) with HLOD scores over 3.00. This was observed on chromosome 18, from marker rs3747899 to marker rs930027 (spanning ~0.5Mb). This region was not previously identified by the microsatellite marker genome-wide scan. Upon further analysis, no individual family produced LOD scores greater than 1.00 and genes in this region were not associated with possible fibrosis/aberrant repair/functionally related pathways involving the lung. Thus it is unlikely that a causal variant in this region is responsible for FPF in all five unrelated families used for this analysis (Kamel, 2010). Due to these factors, this region was not looked at in any further detail.

Finally, the SNP marker genome-wide scan analysis identified markers with LOD scores greater than 1.00 on chromosome 1p in family R0851. This was interesting since the microsatellite marker genome-wide scan also showed suggestive linkage (LOD score ranging from 1.01 to 1.36) in this 1p region. When these findings were combined, chromosome 1p22 was deemed as a good candidate locus (Kamel, 2010) and two candidate genes from this region were picked for sequencing.

### **1.6.6 Genes sequenced**

From the regions of interest identified by microsatellite marker genome-wide scan, microsatellite marker fine mapping and SNP marker genome-wide scan, there were three positive candidate loci containing genes that could potentially have an FPF-causing variant. From these regions of interest, 13 genes were picked as possible candidates based on gene function and lung tissue expression (Kamel, 2010). The largest transcript was



picked for each gene according to Ensembl Assembly GRCh37 and all exons of the gene were fully sequenced.

From the region of interest on chromosome 1, *snoRNA66* and *PINIL* were fully sequenced in at least two FPF patients from all six families (R0851, R0892, R0896, R0942, R1136 and R1487) used for the initial microsatellite marker genome-wide scan. From the ROI on chromosome 6, *TFAP2a*, *TFAP2 $\beta$* , *WRNIP1* and *SERPINB1* were fully sequenced in at least two FPF patients from all six families used for the initial microsatellite marker genome-wide scan. From the region of interests on chromosome 16, *DEXL*, *EMP2*, *VASN*, *GLIS2*, *TIDI1*, *ABCA3*, and *PPL* were fully sequenced in at least two FPF patients from all six families used for the initial microsatellite marker genome-wide scan. All variants found were investigated further and excluded as being FPF-causing by Mr. Fady Kamel (2010).

#### **1.6.7 Telomere assays**

The lengths of telomeres for selected FPF patients from R0851, R0892, R0896, R0942 and R1136 who were still alive were determined by Repeat Diagnostics (British Columbia, Canada). Briefly, white blood cells were isolated from venous blood samples. Repeat Diagnostics measured the lengths of telomeres from lymphocytes and granulocytes by *in situ* hybridization and flow cytometry. The telomere lengths were compared by using a percentile ranking system to average values of telomere lengths seen in a control Caucasian population (Kamel, 2010). In Family R0942, telomere assay analysis on lymphocytes and granulocytes in two affected individuals showed these individuals did not have significantly shortened telomeres (Figure 1.9). From this analysis

it was concluded that the FPF-causing variant is probably not residing in a gene involved in a telomere maintenance pathway, such as *TERT*.

### **1.7 Hypotheses**

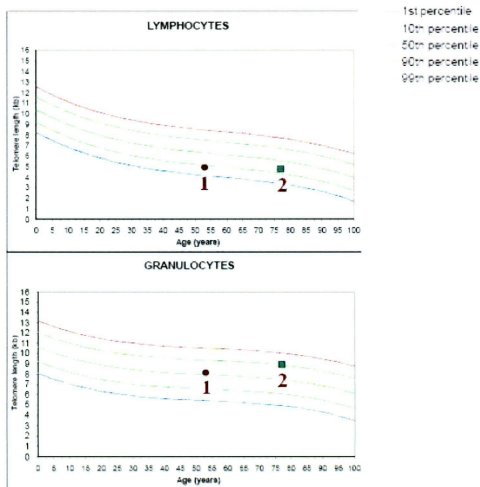
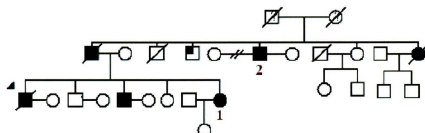
The hypotheses of this project are as follows:

1. There is a novel genetic variant that causes FPF in Family R0942.
2. Newfoundland is an island with a series of founder populations. Thus, there are likely multiple novel genetic variants that cause IPF in the Newfoundland population. Of these, the rs35705950 *MUC5B* promoter variant may be responsible for conferring risk of IPF in the population.

### **1.8 Objectives**

The purpose of this study is two-fold. The first objective is to utilize previous work done to continue the candidate gene sequencing analysis in order to determine the genetic variant(s) that causes FPF in Family R0942. The second objective is to test the significance of a recently discovered *MUC5B* promoter variant and its association with risk of developing IPF.

## R0942



**Figure 1.9: Average telomere lengths for lymphocyte and granulocyte cells in two affected individuals from family R0942 (partial pedigree). Adapted with permission from Kamel, 2010 (Appendix J). Numbers on the graphs correspond to the same numbers on the pedigree. The legend shows the mean values for a control population.**

## **2. Materials and methods**

### **2.1 Study Population**

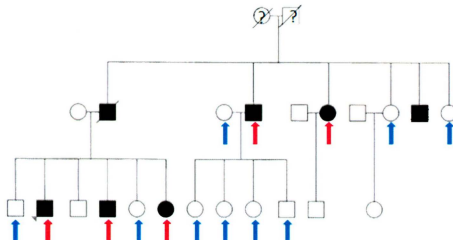
For the purpose of this thesis, a family is described as FPF if there are two or more family members with IPF that are 1<sup>st</sup> or 2<sup>nd</sup> degree relatives (Armanios et al., 2007). Using this criteria, there are 31 families in Newfoundland that have been ascertained for this study. From these 31 families, six families were previously chosen to be investigated further using a candidate gene approach since these families exhibited the greatest potential to find an FPF-causing variant (Kamel, 2010). One (R0942 originating from South Cove, Newfoundland) of these six families was chosen to be investigated for an FPF-causing variant by using a candidate gene approach in this thesis. This family was selected since this family appeared to have the strongest genetic predisposition of all the unsolved families since there are seven affected family members in just two generations, DNA samples were available for all seven affected individuals and ten unaffected family members. Finally, the family pedigree displayed a clear autosomal dominant inheritance pattern (Figure 2.1).

For *MUC5B* rs35705950 SNP genotyping, sporadic IPF cases (n=61) and all familial cases (n=291) with DNA were used. The 291 familial samples consisted of 49 affected and 242 unaffected individuals that were successfully used for genotyping the rs35705950 SNP in *MUC5B*. An additional 277 healthy individuals, previously recruited for a colorectal cancer study by random-digit-dialing (Wang et al., 2009), served as a control cohort for rs35705950 genotyping. All control samples tested were anonymous and could not be linked back to the individual person (Woods et al., 2010). All

participants consented to the use of their DNA for this study. These procedures and the use of the samples were approved by the Health Research Ethics Authority, St. John's, NL, Canada (Study Reference #12.033).

## **2.2 Candidate Gene Selection**

Since one of the focuses of this thesis is on family R0942, the microsatellite marker genome-wide scan, fine mapping, and SNP marker genome-wide scan data was reanalyzed. In total, there were five affected individuals nine unaffected individuals who were genotyped using 550,000 SNPs distributed across the entire genome (Figure 2.1). Since the highest LOD scores in the SNP genome-wide analysis on family R0942 were located on chromosome 6p24.3-6p23 and 16p13.3 (Manhattan plot in Appendix A), the genes in these regions (386 genes, chromosome 16p13.3; 134 genes, chromosome 6p24.3-6p23) were regarded as positional candidates. After determining the regions of interest on chromosomes 6 and 16, a candidate gene table was constructed (Appendix B). All genes in the region of interests, known and hypothetical, were included in the table. The function of the genes was determined by using the websites Online Mendelian Inheritance in Man (OMIM - <http://www.ncbi.nlm.nih.gov/omim>) and UCSC Genome Browser (<http://genome.ucsc.edu>; Kent et al., 2002). Gene expression in different tissues of the human body was determined by the BioGPS website (<http://www.biogps.org/>), a free online gene annotation portal (Wu et al., 2009). Furthermore, protein interaction data was also gathered using the STRING database of known and predicted protein interactions (<http://string-db.org/>).



**Figure 2.1: Pedigree of family R0942 with individuals genotyped for SNP-genome wide scan analysis: 5 affected individuals genotyped (red), 9 unaffected individuals genotyped (blue).**

Gene functions, expression in adult tissues, interesting protein interactions for each gene and any previously known associations with lung disease were all put into the candidate gene table. Using this combined information, a somewhat arbitrary scale for prioritizing which genes to sequence first was determined. The likelihood of each gene harbouring a potential FPF-causing variant in family R0942 was ranked using the following scale from most likely candidate to most unlikely candidate: possible > possible but unlikely > unlikely > very unlikely. All of the ‘possible’ candidate genes were numerically ranked to determine priority for analysis by DNA sequencing.

### 2.3 Oligonucleotide Primer Design

Primers (both forward and reverse directions) were designed using the Primer3 v.0.4.0 webtool (<http://frodo.wi.mit.edu/primer3/>). Primer3 is a useful tool since

important parameters such as primer length, melting temperature, GC content, binding location and likelihood of primer-dimer formation can be noted (Rozen and Skaletsky, 2000). Primer pairs designed ideally were between 18-22bp in length, had melting points within 1°C of each other and were no greater than 65% GC content. Also, primers were designed to be at least 20bp away from the start/end of the target sequence. Once the ideal primers were determined, primer sequences were imported into the BLAT Genome Search tool (<http://genome.ucsc.edu/cgi-bin/hgBlat?command=start>) to confirm each primer was specific for that particular region only. Primer sequences were labeled and ordered from Eurofins (Alabama, USA). See Appendix C for primer sequences and labels. Once arrived, primers were centrifuged for a few seconds and diluted to 1µg/µl by adding the appropriate amount of UltraPure DNase and RNase free distilled water from Life Technologies (Ontario, Canada). Furthermore, a 1:10 dilution was made using 10µl of the initial primer dilution and 90µl of distilled water (dH<sub>2</sub>O). The 1:10 dilutions of primers [100ng/µl] were then ready for use in the PCR reaction.

The reference sequence used in comparison to the experimental sequence was obtained from Ensembl ([http://useast.ensembl.org/Homo\\_sapiens/Info/Index](http://useast.ensembl.org/Homo_sapiens/Info/Index)). The reference sequence used was the longest protein-coding transcript of the gene in question- this provided the most encompassing coverage for analysis of genetics variations.

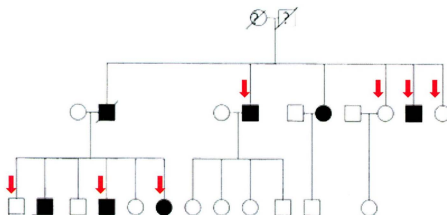
## **2.4 Polymerase Chain Reaction / DNA Sequencing**

Polymerase Chain Reaction (PCR) is a widely-used method for the synthesis/amplification of specific DNA sequences. It utilizes two oligonucleotide primers to hybridize to a region of interest on opposite sides of the double-stranded DNA

(Erich, 1989). After hybridization of a primer with complementary sequence, a DNA polymerase and free deoxyribonucleotide triphosphates (dNTPs) are utilized to copy the template strand. The product can be used as a template for the next cycle, thus the DNA region of interest can theoretically double every cycle- this is done by repeating cycles of denaturation, primer annealing and extension (Erich, 1989). Specifically, the following steps were followed to perform a standard PCR reaction:

A mastermix consisting of 1.5 $\mu$ l of 10X PCR Reaction Buffer, 0.75 $\mu$ l of 50mM MgCl, 0.375 $\mu$ l of dNTPs, 0.15 $\mu$ l of Platinum *Taq* Polymerase, 0.50 $\mu$ l of forward primer, 0.50 $\mu$ l of reverse primer and 9.475 $\mu$ l of dH<sub>2</sub>O was made up. In some instances, 0.75 $\mu$ l of Dimethyl sulfoxide (DMSO) was also added. The amounts of each material used was multiplied by the number of reactions carried out x 1.1 (i.e. if 20 DNA samples were being tested: each of the amounts of PCR Reaction Buffer, 50mM MgCl, dNTPs, Platinum *Taq* Polymerase, forward primer, reverse primer, dH<sub>2</sub>O and in some cases DMSO would be multiplied by 22) and this would be added to the mastermix. All materials except primers were obtained from Life Technologies (Ontario, Canada). 13 $\mu$ l of the mastermix was aliquoted to each reaction well in a 96-well 0.2ml PCR plate. To each well, except the “negative control(s)”, gDNA from affected or unaffected members of family R0942, either 100ng/ $\mu$ l or 50ng/ $\mu$ l, was added. Genomic DNA from four affected individuals and three unaffected individuals in R0942 were used (Figure 2.2). No genomic DNA was added to the negative control well(s). The PCR plate was capped, vortexed and centrifuged up to 1000rpm. The PCR plate was placed in either a Biometra or Eppendorf thermocycler and a specific thermocycler program was run (Appendix D).





**Figure 2.2: Family R0942 with DNA samples used for Sanger sequencing (in red).**

In order to test the success of PCR amplification for the region of interest, a gel electrophoresis technique was used. Briefly, one gram of Ultrapure Agarose and 50ml of 1X TBE Buffer was mixed and heated by microwave to dissolve the agarose in solution. 3.75µl of SYBR Safe DNA Stain Gel (Life Technologies, Ontario, Canada) was added and the mixture was allowed to set in a gel electrophoresis tray. Once solidified, a mixture containing 3µl each of dye and PCR product was pipetted into each well. In one well, a mixture containing 1µl of 100bp DNA ladder (Life Technologies, Ontario, Canada) and 3µl of dye was added. A PowerPac 300 from Bio-Rad (Ontario, Canada) was used to supply 120V for 35 minutes for gel electrophoresis. An image of the gel was taken using a Kodak UV imager. PCR was considered successful if a band(s) of appropriate length in base pairs was visible under UV light when compared to the DNA ladder with no negative control band(s).

From each PCR product, 8µl was pipetted into individual wells in another 96-well 0.2ml PCR plate. Into each well, 0.5µl of Shrimp Alkaline Phosphatase (1U/µl), 0.5µl of Exonuclease I (10U/µl) and 7.5µl of dH<sub>2</sub>O were added. The PCR plate was capped, vortexed and centrifuged up to 1000rpm. The PCR plate was placed in either a Biometra or Eppendorf thermocycler and the “Exosap” program was run (Appendix D).

From this “exosap” product, 4.0µl was pipetted into a Microamp Optical 96-well reaction plate. To each well, the following was added: 0.5µl of BigDye Terminator v3.1 Cycle Sequencing Mix, 2.0µl BigDye Terminator v3.1 5x Sequencing Buffer (Applied Biosystems, Ontario, Canada), 13.83µl of dH<sub>2</sub>O and 0.67µl of primer (forward). The same reagents were added to different wells for a second time with the only change being 0.67µl of reverse primer used during the second time. The Microamp reaction plate was capped, vortexed and centrifuged. The Microamp reaction plate was placed in either a Biometra or Eppendorf thermocycler and the “Abiseq” program was run (Appendix D).

To precipitate the desired product for sequencing, 5µl of 125mM Ethylenediaminetetraacetic acid (EDTA) and 65µl of 100% ethanol (EtOH) was added to every reaction well. The reaction plate was re-capped, vortexed and stored in a drawer (in the dark) for a minimum of 2 hours to a maximum of 24 hours in order to precipitate the desired product for sequencing.

To continue EtOH precipitation of cycle sequencing reactions, the plate was centrifuged for 30 minutes at 3000g. The caps were then removed and the plate was inverted onto clean, folded paper towels to remove the EtOH. The inverted plate and paper towel were placed in the centrifuge for a quick spin at 200rpm. 150µl of 70% EtOH was added to every well. The caps were placed back on the plate and the plate was

centrifuged for 15 minutes at 3000g. After this period, the caps were removed and the plates were inverted again onto clean, folded paper towels. The inverted plate and folded paper towels were placed in the centrifuge for a quick spin at 200rpm. 15 $\mu$ l of HiDi formamide was added to every sample in the plate. The plate was capped, vortexed and centrifuged for a quick spin at 600rpm. The plate was placed onto a thermocycler and the program “denature” was run (Appendix D). After the denaturing thermocycler process, caps were removed and a Septa sequencing mat was fitted onto the Microamp reaction plate. The plate was placed in a reaction cover and put in the ABI prism 3130 Automated Genetic Analyzer for automated sequencing to proceed.

Raw sequencing data was extracted from the ABI prism 3130 and quality of sequence was assessed automatically by the Sequencing Analysis 5.2 program and manually (Kamel, 2010) as well. Sequence of acceptable quality (bi-directional coverage of all exonic bases) was assessed manually using the Sequencher 4.9 or 5.0 programs.

## **2.5 Genomic Regions Analyzed / Characterizing Variants**

Complete bi-directional sequencing was done for all exons (coding and non-coding) in all candidate genes analyzed. The sequences were independently verified for quality control. Furthermore, 25bp of each intron-exon boundary was analyzed. This is because the intronic regions were not always covered at distances greater than 50bp from the exon start locations. Also, since ~50bp away from the exon is typically where the primers were designed to bind, immediate sequence next to the primer binding site was of poor quality. Due to this 25bp cutoff, variants found outside of these cutoffs were reported but not analyzed further for any possible effects.

All variants found within exons and within 25bp of intron-exon boundaries were further analyzed by a systematic method. First, a search was conducted using the dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP/>) and Ensembl (<http://useast.ensembl.org/index.html>) websites to determine if the variant was previously reported in the literature. Based on this search, a variant was labeled as 'previously reported' or 'novel'. If the variant was 'previously reported', the Caucasian population frequency was noted since all of the Newfoundland families in this study were Caucasian. Variants were investigated further if the Caucasian population frequency was <1% and/or if variants segregated with the disease (since FPF is a rare autosomal disease, the expected population frequency is expected to be very low). Reported effects of the protein were noted and further bioinformatic testing was done to determine the potential importance of the variant allele. For all variants that were 'novel' and/or did not have Caucasian frequencies listed, segregation of the variant in affected and unaffected individuals in family R0942 was determined. If the variant did not segregate with the disease, it was ruled out as being causative (shown in results Tables 3.2 and 3.3). The SNPs found in these candidate genes were not phased.

The bioinformatics tools/software utilized included: PolyPhen (Polymorphism Phenotyping, <http://genetics.bwh.harvard.edu/pph/>), a tool "which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations" (Ramensky et al., 2002); Panther (Protein Analysis Through Evolutionary Relationships, <http://www.pantherdb.org/>), a tool designed by expert biologists that classifies genes by their functions using evolutionary relationships and experimental evidence; and SIFT (Sorting Intolerant from

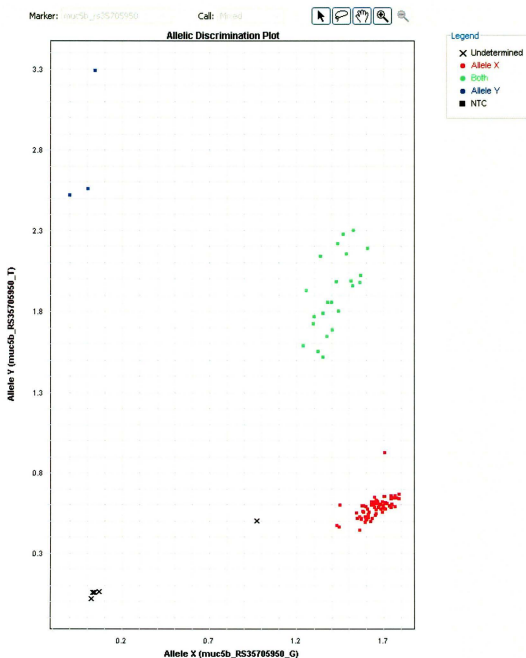
Tolerant, <http://sift.jcvi.org/>), a tool that predicts if an amino acid substitution affects protein function and what impact this change may have.

## **2.6 *MUC5B* rs35705950 SNP genotyping**

### **2.6.1 TaqMan genotyping procedure**

To analyze the frequency and segregation of the *MUC5B* rs35705950 SNP in our study population, a 40X TaqMan SNP Genotyping assay mix (Assay ID: C-1582254-20, part number 4351379) was ordered (Applied Biosystems, California, USA). This kit contained specific probes and reagents required for genotyping the rs35705950 SNP. The kit was stored at -20°C. To prepare the reaction, a mastermix consisting of 525µl 2X TaqMan Universal PCR Master mix, 13.125µl of 40X TaqMan SNP Genotyping assay mix, 13.125µl of TE buffer (consisting of tris-hydroxymethyl-aminomethane and EDTA) and 393.75µl of dH<sub>2</sub>O was made. 9µl of the mastermix was aliquoted into each well of a 0.1ml MicroAmp Fast Optical 96 reaction well PCR plate. 1µl of [50ng/µl] genomic DNA was also added into each reaction well, bringing the total volume to 10µl per reaction well. A MicroAmp Optical Adhesive Film was applied to the PCR plate and the plate was then centrifuged at 1000rpm for approximately 5 seconds.

To perform SNP genotype analysis, the experimental PCR plate was placed in an Applied Biosystems 7900HT Real-time PCR analyzer. Sequence Detection Systems (SDS) v2.4 software was used to construct Allelic Discrimination (AD) and Allelic Quantification (AQ) curves (Appendix E). The 7900HT performed a pre-read in the AD file, an AQ run, and a post-read in the AD file (Appendix E). The data from these reads were automatically called by SDS and a genotype plot was obtained (Figure 2.3).



**Figure 2.3: SDS Allelic Discrimination genotype plot for *MUC5B* rs35705950**  
 Sample of a 96 reaction plate. All individual coloured dots and X's represent individual genotype calls. Red and blue dots are homozygous for G/G and T/T respectively. Green dots are heterozygous. X's consist of 3 control reactions with no DNA and undetermined DNA samples.

Some samples failed to genotype probably due to degradation of DNA, quantity and/or quality of DNA, or contamination of the PCR well. All experimental samples (6.0%) that failed to be genotyped by the TaqMan assay were repeated again. If a sample failed to be genotyped twice, it was excluded from further analysis. The genotype plot was saved and genotype data for all samples was exported and analyzed.

To confirm that genotype calls made by the 7900HT Real-time PCR analyzer were correct, a subset of samples that were successfully genotyped (n=20) was randomly selected. These 20 samples consisted of eight G/G genotype samples, eight G/T genotype samples and four T/T genotype samples. Primers were designed to cover the *MUC5B* promoter SNP (Primer sequences in Appendix C). PCR and Sanger sequencing were performed to validate if the calls were correctly made. All 20 calls made by the 7900HT Real-time PCR analyzer in this subset were validated, confirming the accuracy of calls made by this system.

### **2.6.2 Description of clinical variable parameters**

Clinical variables were also gathered for all of the cases, (n=110) and controls (n=277) used in genotyping. Cases consisted of sporadic IPF individuals (n=68) and familial PF individuals (n=48) from 12 affected families. Variables such as age, gender, smoking status, smoking duration, pulmonary function tests and method of diagnosis were all gathered from electronic clinical records by Mrs. Barbara Noble (Table 3.4). For the controls, clinical variables are based on n=270 since questionnaires were not returned by seven participants. These controls were from a colorectal cancer study in Newfoundland (Woods et al., 2010) and no pulmonary function information was gathered.

Age and smoking status were variables that had sub-classifications. For all cases, age of diagnosis was recorded. For all controls, age of sample collection was recorded. Smoking status data was divided into four categories: smokers, ex-smokers, occasional smokers and never smoked. Smokers were individuals who were actively smoking at the time of diagnosis and/or sample collection. Ex-smokers were individuals who had stopped smoking at the time of diagnosis and/or sample collection. Out of all cases, there were 73 ex-smokers. 71.2% of these individuals had smoked for over 20 years. Out of all control individuals, 128 were ex-smokers. 53.1% of these individuals had smoked for over 20 years. Occasional smokers were classified as individuals who do not smoke habitually and only smoke in certain social settings. Out of the five total occasional smokers in the study, only one had smoked occasionally for over 20 years.

### **2.6.3 *MUC5B* case-control statistics**

In order to assess the association of the rs35705950 *MUC5B* promoter polymorphism with IPF in the Newfoundland population, a case-control study was carried out utilizing various statistical methods. To determine Hardy-Weinberg equilibrium (HWE) of genotype frequencies for cases and controls, a Chi-squared Hardy-Weinberg equilibrium test calculation for biallelic markers was used (Rodriguez et al., 2009). Chi-square values ( $\chi^2$ ) represent the difference between observed and expected genotype frequencies. If the P value > .05 then genotype frequencies are consistent with HWE and what is expected and if P < .05, the genotype frequencies are significantly different than what is expected.

To determine if there is a significant difference between cases and controls for rs35705950, a logistic regression was carried out. Logistic regression was used because it



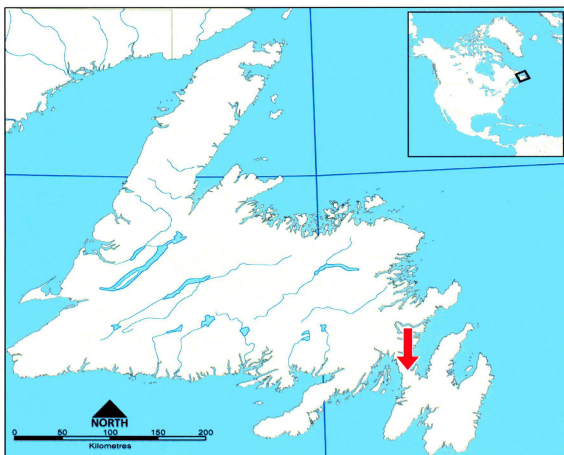
is a generalized linear model that can predict the probability of an outcome (i.e. disease) based on one or more predictor values (i.e. genotype) by yielding odds ratio and confidence interval values. If the odds ratio is greater than one, the outcome is more likely to occur than not. In our case, the outcome was the disease phenotype. Therefore, if the odds ratio was greater than one, risk of PF was more likely than not. Conversely, if the odds ratio was less than one, this would indicate a preventative effect rather than a risk effect. The confidence interval (CI) associated with the odds ratio demonstrates the reliability of the odds ratio number and show the upper and lower boundary of possible effect sizes (Davies, 1998). If a 95% CI includes the odds ratio value that shows an effect, this finding is non-significant at the 5% level. Furthermore, the CI upper and lower limits show how big or small the effect could be (Davies, 1998) but this also depends on the sample size of the study.

The logistic regression type used in this study was multinomial to demonstrate if there was additional predicted risk based on the number of variant alleles possessed. This produced odds ratio/CI for each variant genotype in comparison to the wild-type genotype. Multinomial logistic regression analysis was carried out comparing probands to controls. Furthermore, in order to compare all cases to controls, multinomial logistic regression analysis using familial clusters was done since 27.2% of all cases were related and had to be corrected for. Individuals from a related family were put in the same cluster and all unrelated individuals were placed in different clusters. Statistical correction for related individuals was implemented by using this method.

### **3. Results**

#### **3.1 Clinical description and findings - Family R0942**

Family R0942 is from a small community in Trinity Bay, Newfoundland, Canada. According to the Statistics Canada Census of 2006, the population of this community is slightly over 300. There are other families with FPF in other small communities across the island of Newfoundland, suggesting a series of founder effects leading to the higher-than-average prevalence of FPF.

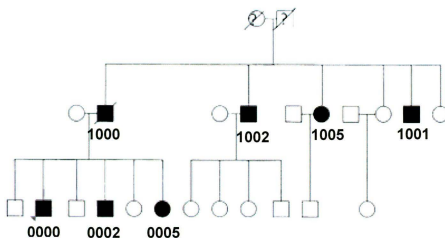


**Figure 3.1: Location of Trinity Bay on a map of Newfoundland.** Location of Trinity Bay is denoted by the red arrowhead.

The clinical histories of individuals in this family have been well-documented by the Provincial Medical Genetics Program (PMGP). The proband of this family is a male (R0942.0000, Figure 3.2) who was diagnosed at 39 years of age. He presented with shortness of breath and pneumonia. Review of the questionnaire he provided showed that he did have some exposure to asbestos and wood. Furthermore, he was a smoker for over 20 years. PFTs, HRCT scan and a lung biopsy were performed. A diagnosis of IPF was given. The proband also stated that his father (R0942.1000) had died of PF and there were others in his nuclear family that had similar symptoms to his, leading to screening of other family members.

After careful screening, IPF was diagnosed in one brother (R0942.0002) and one sister (R0942.0005) of the proband. The brother presented with a chronic cough, shortness of breath and chest pain for many years. He was also a smoker for over 25 years. He was diagnosed by PFTs and HRCT scan at age 46. The sister of the proband presented with shortness of breath and chest pain. She had recently given up smoking but had smoked for 31 years. A diagnosis of IPF was given after PFTs, HRCT scan and a lung biopsy. The proband has three other siblings (two brothers, one sister) who have been screened clinically but have not presented with any symptoms.

The father of the proband (R0942.1000) who had died of IPF also had siblings who were affected. From his six siblings (three brothers, three sisters), two brothers (R0942.1001, R0942.1002) and one sister (R0942.1005) were diagnosed by IPF from screening of the family (Figure 3.2). All three of these siblings had smoked for over 40 years. The other unaffected siblings were also clinically screened but there have been no presenting symptoms yet.



**Figure 3.2: Family R0942 with identifying numbers for affected individuals only.**

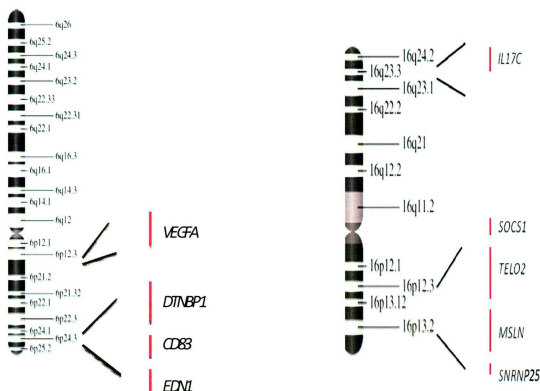
### 3.2 Candidate genes analyzed

Previous microsatellite genome-wide scans, SNP genome-wide scans and microsatellite fine-mapping yielded regions of LD on chromosome 16p and 6, these regions were the focus of candidate gene analysis (Edwards, unpublished data; Kamel, 2010). Mr. Fady Kamel fully sequenced all exons in four genes on chromosome 6 (*TFAP2a*, *TFAP2β*, *WRNIP1* and *SERPINB1*) and six genes on chromosome 16 (*DEXI*, *EMP2*, *VASN*, *GLIS2*, *TID1*, and *PPL*) in two affected individuals in R0942. Sequencing of these 10 genes was unsuccessful in determining a FPF-causing genetic variant(s).

Therefore, additional candidate genes from these two ROI were methodically selected for sequencing. In total, four additional genes from chromosome 6 (*EDN1*, *CD83*, *DTNBP1*, *VEGFA*) and five additional genes from chromosome 16 (*SOCS1*, *TELO2*, *SNRNP25*, *MSLN*, *IL17C*) were fully sequenced (Table 3.1 and Figure 3.3).

**Table 3.1: All genes fully sequenced in the regions of interest on chromosome 6 and 16 from this project**

	Sequenced in 5 affected and 2 unaffected individuals in Family R0942	
Chromosome	6	16
Genes sequenced	<i>EDN1</i> <i>CD83</i> <i>DTNBP1</i> <i>VEGFA</i>	<i>SOCS1</i> <i>TELO2</i> <i>SNRNP25</i> <i>MSLN</i> <i>IL17C</i>



**Figure 3.3: Physical location of genes sequenced within regions of interest on chromosomes 6 and 16.**

These genes were fully sequenced in four affected and three unaffected individuals from family R0942 (Figure 2.2). The unaffected individuals were sequenced to perform a preliminary segregation analysis in case a potential causative variant was found. Two of the unaffected individuals are 78 and 81 years of age and have smoked for 29 and 45 years respectively. It is therefore less likely that they will develop PF based on genetic causes at such an advanced age. The third unaffected individual was used to see if a potential variant segregated with PF in the sibship. All variants found within these seven individuals are reported in Table 3.2 and Table 3.3.

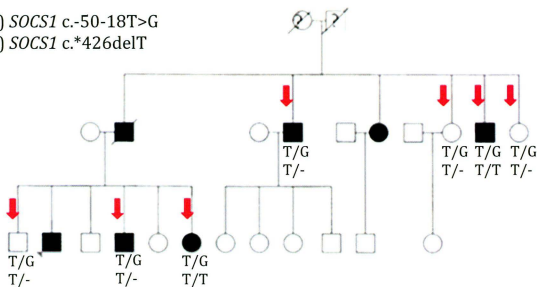
### **3.2.1 Results from candidate gene sequencing - Chromosome 16**

#### **3.2.1.1 *SOC***

*SOC* is a gene located on chromosome 16: 11348274-11350039 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). There is only one transcript for *SOC* which is made up of two exons consisting of a total of 1225 base pairs and codes for a 211 amino acid protein. *SOC* is a member of the suppressor of cytokine signaling family of genes. Cytokine and cytokine signaling proteins have previously been demonstrated to be involved in molecular pathways leading to PF and are thus of interest. A study also indicates *SOC* is a suppressor for pulmonary fibrosis since there was a significantly downregulated expression of *SOC* in the lungs of IPF patients (Nakashima et al., 2008). Furthermore in this study, viral gene therapy with *SOC* in mice reduced inflammation and PF, suggesting its importance in the pathogenesis of PF (Nakashima et al., 2008). Even though *SOC* gene expression results were uninteresting (Figure 3.8), the functional significance of *SOC* in relation to PF made this an ideal candidate gene for this study.

Both exons of *SOCS1* were fully sequenced in the seven selected individuals from R0942. Two variants were found, both of which were previously reported variants (Table 3.2). One variant was a substitution located in intron 1 (c.-50-18T>G). This was labeled non-pathogenic since the frequency in the HapMap Caucasian population was 0.992 and it did not segregate with FPF in this family. The other variant found was a deletion in the 3' UTR (c.\*426delT). This variant was also excluded as FPF-causing because it did not segregate with FPF in this family (Figure 3.4). Furthermore, it was previously reported in other control populations.

- 1) *SOCS1* c.-50-18T>G
- 2) *SOCS1* c.\*426delT



**Figure 3.4: Segregation analysis of both variants found in *SOCS1*.** Segregation of *SOCS1* c.-50-18T>G is shown in the first line and segregation of *SOCS1* c.\*426delT is shown in the second line. Red arrowheads indicate individuals with DNA sequenced. These variants did not segregate with the disease in this family.

#### 3.2.1.2 *MSLN*

The *MSLN* gene is located on chromosome 16: 810765-818865 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). There are ten protein-coding

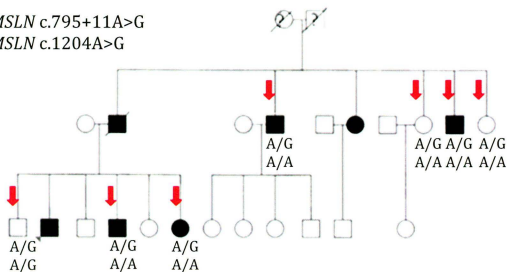
transcripts for *MSLN* with the longest one containing 17 exons and coding for a 630 amino acid protein. According to the BioGPS gene expression profile, *MSLN* is significantly expressed in the lung in comparison to all other tissues in humans. The precursor protein encoded by this gene is cleaved into two products: mesothelin and megakaryocyte potentiating factor (UCSC Genome Browser; Kent et al., 2002). These products are involved in developing the mesothelium, a layer of cells that provides a protective covering for some human organs. Normally, the site of normal developing mesothelium is the lung- specifically its internal chest wall. However, if the mesothelial cells are differentiated abnormally, malignant mesothelioma (MM) can occur. Just like pulmonary fibrosis caused by inhalational agents, MM is most commonly caused by asbestos exposure (Tan et al., 2010). Therefore it is possible there are some overlapping molecular mechanisms in both diseases' pathogenesis. Along with a significantly higher expression in the lung, these observations led to *MSLN* being selected as a candidate gene.

All 17 exons of the longest transcript of *MSLN* were fully sequenced in the seven selected individuals from R0942. Three variants were found in this gene of which two were intronic. One of the intronic variants was found in introns 8 (c.795+11A>G). This previously reported SNP was present in all seven affected and unaffected individuals sequenced in R0942 and thus did not segregate with FPF (Figure 3.5). Also, the allele frequency was 0.967 based on a pilot-1 CEU population, indicating it is present at a high frequency in the Caucasian population. The other intronic variant (c.1230+16G>T) was located in intron 11 and had a 1000 Genomes minor allele frequency of 0.018. However, this SNP was previously reported in the literature, did not segregate with affected



individuals in R0942 and had no predicted impact on splicing or protein function. Hence this variant was excluded as a potential FPF-causing variant. The only SNP found in an exon of *MSLN* was located in exon 11 (c.1204A>G). This substitution caused a protein change, p.Asn402Asp. Since this SNP was only present in one affected individual in R0942, it was also excluded as a potential FPF-causing variant in this family (Figure 3.5). Furthermore, the 1000 Genomes minor allele frequency was 0.095.

- 1) *MSLN* c.795+11A>G
- 2) *MSLN* c.1204A>G



**Figure 3.5: Segregation analysis of three variants found in *MSLN*.** Segregation of *MSLN* c.795+11A>G is shown in the first line and segregation of *MSLN* c.1204A>G is shown in the second line. Red arrowheads indicate individuals with DNA sequenced. These variants did not segregate with the disease in this family.

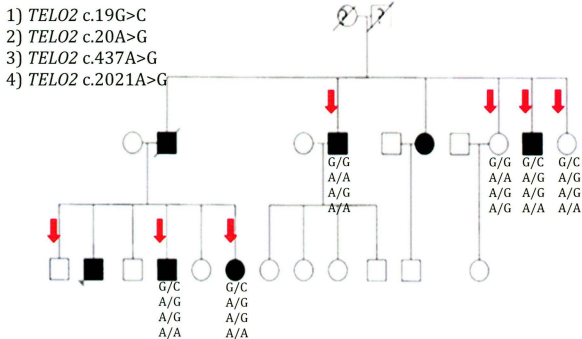
### 3.2.1.3 *TELO2*

*TELO2* is a gene located on chromosome 16: 1543352-1560460 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). There are three protein coding transcripts for this gene. The longest transcript contains 21 exons and codes for an 837 amino acid protein. The functions of *TELO2* are largely unknown in mammals but it is

suggested to be involved in the regulation of telomere length. This is of importance since variants in *TERT*, a gene involved in preventing the shortening of telomeres, are the most common documented genetic cause of FPF (Tsakiri et al., 2007; Alder et al., 2008; Armanios et al., 2007; de Leon et al., 2010). Furthermore, it is possible that *TELO2* has additional functions besides that of telomere maintenance since it is highly conserved in *Drosophila*. *Drosophila* and other insects of the same order lack the canonical telomerase-based telomere system thus the conservation of *TELO2* sequence in this order is interesting (Takai et al., 2007). *TELO2* also has median expression in the lung and interacts with other proteins involved in DNA repair pathways (Figure 3.9). Due to its involvement in telomere maintenance and other possible functional involvements, *TELO2* was selected as a candidate gene.

All 21 exons of *TELO2* were fully sequenced in the seven selected individuals from R0942. Eight variants were found in this gene, all of which were previously reported (Table 3.2). Three of the eight variants were intronic substitutions (c.682+115C>T; c.2226+38A>G; c.2227-39C>T) and were excluded as FPF-causing because they were located more than 30 bases into the introns, were all present at an allele frequency of greater than 1%, had no predicted impact on protein or gene expression and did not segregate with pulmonary fibrosis in this family. There were two variants found in exon 21: one variant in exon 21 (c.\*174C>T) was found and labeled non-pathogenic since it was located in the non-coding part of the exon and had a HapMap CEU allele frequency of 0.367. Another substitution variant (c.2415T>C) was labeled non-pathogenic since the substitution resulted in a synonymous amino acid change and the allele frequency in the HapMap CEU population was 0.429. Three variants (exon 2: c.19G>C, c.20A>G; exon 3:

c.437A>G; exon 16: c.2021A>G) resulted in a change in amino acid. However, these variants did not segregate with FPF in this family and thus were labeled as non-disease causing (Figure 3.6).



**Figure 3.6: Segregation analysis of three missense variants in *TELO2*.**

Segregation of *TELO2* variants c.19G>C, c.20A>G, c.437A>G and c.2021A>G are shown in the first, second, third and fourth lines respectively. Red arrowheads indicate individuals with DNA sequenced. These variants did not segregate with the disease in this family.

#### 3.2.1.4 *SNRNP25*

The *SNRNP25* gene is located on chromosome 16: 103829-107669 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). *SNRNP25* has three protein coding transcripts. The largest one contains five exons and codes for a 132 amino acid protein. This gene encodes a small nuclear ribonucleoprotein (SNRNP) with a

molecular weight of 25 kDa. SNRNPs are made up of small nuclear RNAs and are involved in cellular processes such as splicing of RNA transcripts to make mRNAs. *SNRNP25* encodes a protein that is a component of the U12-type spliceosome which functions in removing introns with a unique splice sequence for recognition (UCSC Genome Browser; Kent et al., 2002). Recently, an additional role for *SNRNP25* was discovered. A study showed that the prevalence of PF was the key significant difference between patients with or without anti-U11/U12 antibodies (Fertig et al., 2009). The results of this study concluded that presence of anti-U11/U12 antibodies is a novel biomarker for PF (Fertig et al., 2009). Following this, it is postulated that a genetic change in one of the genes coding for the proteins of the U12-type spliceosome could alter the structure of the spliceosome and thus cause the presence of these specific antibodies in PF patients. Therefore, *SNRNP25* was selected as a candidate gene.

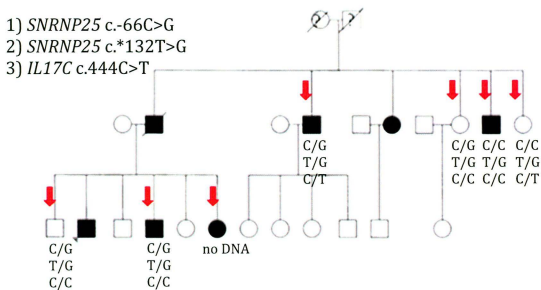
All five exons of *SNRNP25* were fully sequenced in the seven selected individuals from R0942. There were three substitution variants found in this gene (Table 3.2). One variant (c.-66C>G) was located in the 5'UTR; the second variant found (c.70-15C>T) was located in intron 1; the last variant found (c.\*132T>G) was located in the 3'UTR. All three variants were previously reported in the literature, were of unknown predicted impact, did not segregate with FPF in R0942 (Figure 3.7) and had a greater minor allele frequency than 0.01 in dbSNP. Due to these observations, all three variants found in *SNRNP25* were excluded as potentially pathogenic genetic variants.

### 3.2.1.5 *IL17C*

*IL17C* is located on chromosome 16: 88705001-88706882 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). This gene has one protein-coding

transcript which contains three exons and codes for a 197 amino acid protein. *IL17C* encodes a cytokine that is involved in releasing tumor necrosis factor alpha and other interleukins from activated cells (UCSC Genome Browser; Kent et al., 2002). The IL17 family of cytokines has been documented to be involved in many pulmonary inflammatory responses (Kawaguchi et al., 2004). The cytokine pathway leading to pulmonary inflammation and ECM remodeling is a major contributor to the pathogenesis of IPF. Furthermore, cytokines from the IL17 family can induce the expression of *MUC5B* (a known IPF-causing gene) in bronchial epithelial cells, lending further support to its potential involvement in an IPF molecular pathway (Kawaguchi et al., 2004). From these functional considerations, *IL17C* was selected as a candidate gene.

All three exons of *IL17C* were fully sequenced in the seven selected individuals from R0942. There was only one variant found in this gene (Table 3.2). This variant was a substitution (c.444C>T) that resulted in a synonymous change (p.Ser148=). The variant was previously reported in the literature, had a minor allele frequency of 0.041 (1000 Genomes Project) and did not segregate with FPF in R0942 (Figure 3.7). Thus, this variant was classified as non-pathogenic.



**Figure 3.7: Segregation analysis of a 5'UTR variant and a 3'UTR variant in *SNRNP25* and an exonic variant in *IL17C*.** Segregation of *SNRNP25* variants c.-66C>G and c.\*132T>G are shown in the first and second lines respectively .

Segregation of *IL17C* variant c.444C>T is shown in the third line. Red arrowheads indicate individuals with DNA sequenced. These variants did not segregate with the disease in this family.

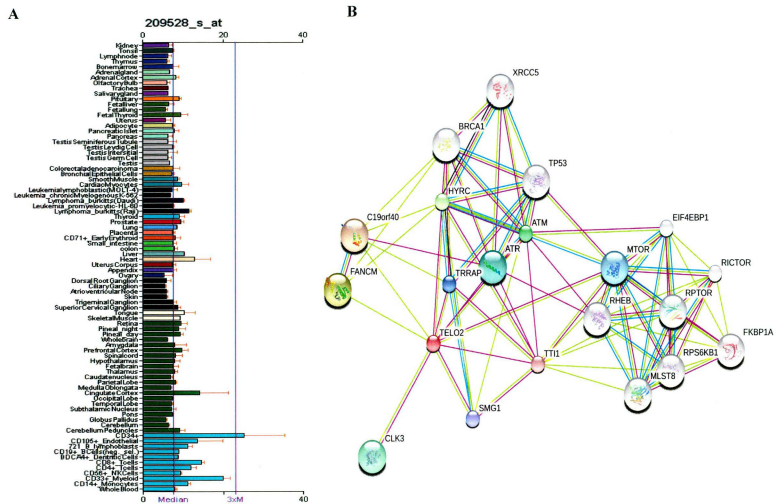
**Table 3.2: Characteristics of all variants found by Sanger sequencing on chromosome 16 ROIs**

Gene	Exon	Variant	Type	Amino acid change	Predicted impact	dbSNP ID	Caucasian Pop. Frequency	Additional information
SOCS 1	intron 1	c.-50-18T>G	substitution	no	Unknown	rs27829	HapMap CEU: 0.992	does not segregate with FPF
	2	c.*426del T	deletion	no	Unknown		n/a	in 3' UTR region, does not segregate with FPF
TELO 2	2	c.19G>C, c.20A>G	2bp substitution	p.Glu7Arg	n/a	rs2667660 rs2667661	n/a	does not segregate with FPF
	3	c.437 A>G	substitution	p.Gln146 Arg	Missense	rs2235624	HapMap CEU: 0.440	>1% population frequency
	intron 4	c.682+11 5C>T	substitution	no	Unknown	rs2235630	pilot 1 CEU: 0.750	>30 bases into the intron
	16	c.2021A>G	substitution	yes	Gln-->Arg	rs2248128	HapMap CEU: 0.372	does not segregate with FPF
	intron 18	c.2226+3 8A>G	substitution	no	Unknown	rs11248882	dbSNP: a vg het: 0.455	>30 bases into the intron
	intron 18	c.2227-39C>T	substitution	no	Unknown	rs3736721	HapMap CEU: 0.250	>30 bases into the intron
	21	c.2415T>C	substitution	p.Ala805 =	Synonymous	rs3180228	HapMap CEU: 0.429	>1% population frequency
	21	c.*174C>T	substitution	no	Unknown	rs9454	HapMap CEU: 0.367	>in non-coding region
	5' UTR	c.-66C>G	substitution	no	Unknown	rs2562145	1000 Genome s MAF: C=0.463/583	>1% population frequency
SNRN P25	intron 1	c.70-15C>T	substitution	no	unknown		HapMap CEU: 0.460	does not segregate with FPF
	3'UTR	c.*132T>G	substitution	no	Unknown	rs1045001	HapMap CEU: 0.301	>1% population frequency
MSLN	intron 8	c.795+11 A>G	substitution	no	unknown	rs1737340	pilot 1 CEU: 0.967	high population frequency
	11	c.1204A>G	substitution	p.Asn402 Asp	Missense	rs73491255	1000 Genome s MAF:	does not segregate with FPF

							0.095	
	intron 11	c.1230+1 6G>T	substitution	no	unknown	rs734912 58	1000 Genome s MAF: 0.018	\does not segregate with FPF
<i>IL17C</i>	3	c.444C>T	substitution	p.Ser148 =	Synonymo us	rs110766 88	1000 Genome s MAF: 0.041	does not segregate with FPF







**Figure 3.9: A) BioGPS human tissue gene expression profile for *TELO2*. B) STRING *TELO2* protein to protein interactions**

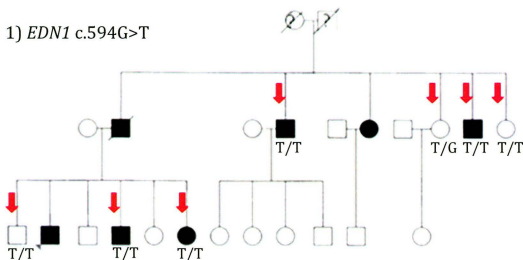
### 3.2.2 Results from candidate gene sequencing - Chromosome 6

#### 3.2.2.1 *EDNI*

The *EDNI* gene is located on chromosome 6: 12290529-12297427 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). This gene has one protein-coding transcript containing five exons and coding for a 212 amino acid protein. *EDNI* encodes the endothelin-1 protein which participates in a variety of physiological pathways in the human body. Specifically in the lung, endothelin-1 drives multiple processes that all lead to ECM deposition such as fibroblast activation, proliferation, and differentiation into myofibroblasts (Swigris and Brown, 2010). The fibroblast pathway that leads to excess ECM deposition is one of the hallmarks of IPF. Patients with IPF have significantly increased levels of *EDNI* protein in comparison to healthy controls in a number of studies. In a possible-disease model involving *EDNI*, injured AECs would recruit a variety of pro-fibrotic cytokines such as endothelin-1. In turn, aberrant activation and differentiation of fibroblasts to myofibroblasts, excess ECM and collagen deposits and abnormal tissue repair would occur, leading to IPF (Swigris and Brown, 2010). Furthermore, *EDNI* is expressed in lung tissue over 30x more than median in all other tissues, suggesting its profound role in various lung processes (Figure 3.13). The significant expression of *EDNI* in lung tissue and previous functional evidence documenting its role in fibroblast cascades led to the selection of *EDNI* as a candidate gene for Sanger sequencing.

All five exons of *EDNI* were fully sequenced in the seven selected individuals from R0942. There were three variants found in this gene (Table 3.3). One of the variants (c.-131delA) was located in the 5'UTR. This deletion was previously reported in the

literature and had a minor allele frequency of 0.718. Therefore this was labeled a non-pathogenic variant. The second variant found in *EDN1* was located in exon 3 and was a substitution (c.318A>G). This was a homozygous change in all seven individuals and the variant was previously reported in the literature, causing a synonymous amino acid change. Due to these observations, the variant was excluded from being a potential pathogenic variant. The final variant was a substitution (c.594G>T) found in exon 5. This substitution was predicted to cause a missense change in the resulting amino acid (p.Lys197Asn). However, this variant had a HapMap CEU minor allele frequency of 0.071 and was only found in one unaffected individual (Figure 3.10). Thus, this was unlikely to be an FPF-causing variant.

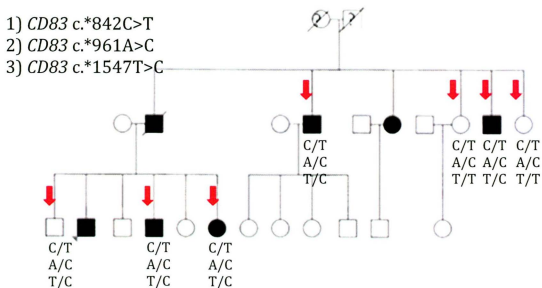


**Figure 3.10: Segregation analysis of an exonic variant found in *EDN1*.** Segregation of *EDN1* variant c.594G>T is shown in the first line. Red arrowheads indicate individuals with DNA sequenced. This variants did not segregate with the disease in this family.

### 3.2.2.2 *CD83*

*CD83* is located on chromosome 5: 14117487-14137148 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). This gene has one protein-coding transcript that contains five exons and codes for a 205 amino acid protein. The *CD83* gene codes for a membrane protein that participates in the regulation of antigen presentation (UCSC Genome Browser; Kent et al., 2002). *CD83* is also expressed by dendritic cells that initiate and control various immune responses in the body (Bantsimba-Malanda et al., 2010). Although there are many dendritic cells that act in the lung, their involvement in IPF has not been documented. A recent study was the first to demonstrate that dendritic cells that express *CD83* accumulate in the lung during bleomycin-induced fibrosis of the lung (Bantsimba-Malanda et al., 2010). Since these cells are involved in a proinflammatory pathway leading to fibrosis, *CD83* was selected as a candidate gene.

All five exons of *CD83* were fully sequenced in the seven selected individuals from R0942. There were five variants found in this gene (Table 3.3). One variant (c.38-43C>G) was located in intron 1 and was labeled non-pathogenic due to its location (43bp before the start of exon 2) and a minor allele frequency of 0.266 in control populations. Three variants were located in the non-coding region of exon 5. These were previously reported substitution variants that were also labeled non-pathogenic since they did not segregate with FPF (Figure 3.11), had a >1% population frequency and had no known functional impact reported by the literature. The last variant found was a substitution (c.207G>A) located in exon 3. This was a synonymous variant that had a minor allele frequency of 0.265 (Table 3.3). This variant was also classified as non-pathogenic.



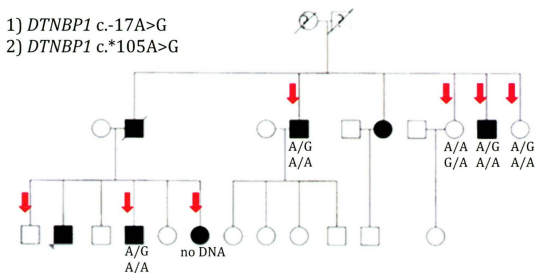
**Figure 3.11: Segregation analysis of three variants in exon 5 of *CD83*.** Segregation of *CD83* variants c.\*842C>T, c.\*961A>C and c.\*1547T>C are shown in the first, second and third lines respectively. Red arrowheads indicate individuals with DNA sequenced. These variants did not segregate with the disease in this family.

### 3.2.2.3 *DTNBP1*

The *DTNBP1* gene is located on chromosome 6: 15523032-15663289 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). The longest protein-coding transcript for this gene contains ten exons and codes for a 351 amino acid protein. *DTNBP1* codes for dystrobrevin binding protein 1 which is expressed ubiquitously and has functional roles in organelle biogenesis and trafficking (Kent et al., 2002). There is evidence showing that a nonsense mutation in *DTNBP1* causes Hermansky-Pudlak syndrome (HPS) type 7 (Li et al., 2003). HPS is a genetically heterogeneous disorder characterized by oculocutaneous albinism, prolonged bleeding and sometimes pulmonary fibrosis (Li et al., 2003). There are seven subtypes of HPS, each with slightly different phenotypical manifestations. In three of the HPS subtypes including HPS type 7, dyspnea

or severe pulmonary fibrosis is also observed (Pierson et al., 2006). Since a nonsense mutation in *DTNBP1* causes HPS type 7, it is plausible that a different genetic mutation could result in PF without the other disease manifestations. From these observations, *DTNBP1* was selected as a candidate gene.

All ten exons of *DTNBP1* were fully sequenced in the seven selected individuals from R0942. There were two variants found in this gene (Table 3.3). One of the variants (c.-17A>G) was a 5' UTR substitution variant and the other (c.\*105A>G) was located in the non-coding region of exon 10. These variants had minor allele frequencies of 0.138 and 0.127 respectively and did not segregate with FPF (Table 3.3; Figure 3.12). These variants were thus classified as non-pathogenic.



**Figure 3.12: Segregation analysis of both variants found in *DTNBP1*.** Segregation of *DTNBP1* variants c.-17A>G and c.\*105A>G are shown in the first and second lines respectively. Red arrowheads indicate individuals with DNA sequenced. These variants did not segregate with the disease in this family.

#### 3.2.2.4 *VEGFA*

*VEGFA* is located on chromosome 6: 43737946-43754223 (UCSC Genome Browser, GRCh37/hg19 assembly; Kent et al., 2002). The longest protein-coding transcript for *VEGFA* contains eight exons and codes for a 412 amino acid protein. *VEGFA* codes for vascular endothelial growth factor a. This is part of a protein complex that targets endothelial cells and precipitates a wide variety of physiological effects such as increasing angiogenesis/ endothelial cell growth, promoting cellular migration and regulating apoptosis. *VEGFA* primarily promotes angiogenesis (blood vessel growth). An imbalance in angiogenic factors such as *VEGFA* and transforming-growth factor  $\beta$  (*TGF $\beta$* ) that leads to abnormal vascular growth has been demonstrated with IPF (Antoniou et al., 2010). Increased vascularization is also displayed around fibroblastic foci and in areas of fibrosis in the lung (Cosgrove et al., 2004). Moreover, expression of *VEGFA* is 10x higher in lung tissue than median value in all other tissues (Figure 3.13). These factors led to *VEGFA* being selected as a positional and functional candidate gene.

All eight exons of *VEGFA* were fully sequenced in the seven selected individuals from R0942. There was one substitution variant (c.534C>T) found in the coding region of exon 1. The C>T substitution resulted in a synonymous amino acid (serine). This variant was previously reported in the literature and had a HapMap CEU minor allele frequency of 0.195 (Table 3.3). Therefore, the variant was classified as non-pathogenic.



**Table 3.3: Characteristics of all variants found by Sanger sequencing on chromosome 6 ROIs**

Gene	Exon	Variant	Type	Amino acid change	Predicted impact	dbSNP ID	Caucasian Pop. Frequency	Additional information
<i>EDN1</i>	5' UTR	[c.-131delA]+[c.-131delA]	deletion	no	unknown	rs10478694	PDR90: 0.718	in 5' UTR region
	3	[c.318A>G]+[c.318A>G]	substitution	no	synonymous: Glu --> Glu	rs5369	HapMap CEU: 0.770	>1% population frequency
	5	c.594G>T	substitution	p.Lys197Asn	missense	rs5370	HapMap CEU: 0.071	does not segregate with FPF
<i>CD83</i>	intron 1	c.38-43C>G	substitution	no	unknown	rs2235368	Pilot 1 CEU: 0.175	>1% population frequency
	3	c.207G>A	substitution	no	synonymous: Arg --> Arg	rs7743206	HapMap CEU: 0.265	does not segregate with FPF
	5	c.*842C>T	substitution	no	unknown	rs1050648	HapMap CEU: 0.150	>1% population frequency
	5	c.*961A>C	substitution	no	unknown	rs1050650	HapMap CEU: 0.460	>1% population frequency
	5	c.*1547T>C	substitution	no	unknown	rs9230	Pilot 1 CEU: 0.242	does not segregate with FPF
<i>DTNBP1</i>	5' UTR	[c.-17A>G]+[c.-17A>G]	substitution	no	unknown	rs11558324	Pilot 1 CEU: 0.192	>1% population frequency
	10	c.*105A>G	substitution	no	unknown	rs1047631	HapMap CEU: 0.271	does not segregate with FPF
<i>VEGFA</i>	1	c.534C>T	substitution	no	synonymous: Ser --> Ser	rs25648	HapMap CEU: 0.195	>1% population frequency



### **3.3 *MUC5B* rs35705950 genotyping**

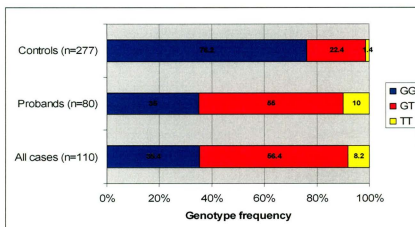
#### **3.3.1 Case vs. control analyses**

A case-control analysis was carried out using 110 affected individuals and 277 healthy controls from the Newfoundland population. From the probands, 30/110 had additional family members with PF. The analysis was performed twice using SPSSv20.0 software with Dr. Guangju Zhai's guidance (Memorial University). This analysis was performed once with all 110 cases while controlling for familial samples via clustering and once by using the 80 unrelated probands without including additional affected family members. The clinical variables/results for both analyses are summarized in Table 3.4.

There was an association between rs35705950 genotypes and IPF (Figure 3.14). The minor allele T was present in 36.4% of affected individuals and 12.6% of healthy controls. The genotype frequencies for controls were consistent with Hardy-Weinberg equilibrium ( $\chi^2 = .05$ ,  $P = .82$ ) but were significantly different for all cases ( $\chi^2 = 5.22$ ,  $P = .02$ ). The odds ratio for individuals affected with IPF who were heterozygous for the variant minor allele was 5.4 (95% confidence interval, 3.3 to 9.6,  $P < .001$ ). The odds ratio for individuals affected with IPF who were homozygous for the variant minor allele was 12.2 (95% confidence interval, 3.3 to 44.7,  $P < .001$ ). Furthermore, there was no significant difference between probands and all cases ( $P > .05$ ).

#### **3.3.2 Familial segregation of *MUC5B* promoter SNP rs35705950**

Since there were 30 additional family members included in one part of the case vs control analyses, a segregation analysis was performed on all 12 families that had been characterized with FPF. From this analysis, rs35705950 was shown to segregate in two families: R1136 and R0942 (Figure 3.15).



**Figure 3.14: Excel bar graph comparing differences between genotype frequencies.** 76.2% of controls had the G/G genotype in comparison to 35% of probands and 35.4% of all cases. Only 23.8% of controls had at least one T allele in comparison to 65% of probands and 64.6% of all cases.

Family R1136 exhibited segregation of the *MUC5B* variant T allele with PF. The proband, diagnosed with PF at 60 years of age, is homozygous for the variant T allele. There are four other affected individuals in this family. Three of these individuals are heterozygous for the variant T allele (diagnosed at ages 53, 61 and 92) and one is homozygous for the variant T allele (diagnosed at age 64). There are 14 other individuals in this family who are heterozygous for the variant T allele and two others who are homozygous for the variant T allele. These 14 individuals are unaffected. In particular, eight unaffected individuals are between 34-48 years old which is before the typical age of onset (~62 years) for FPF in our cohort. All homozygous wild-type individuals were unaffected at last follow up.

Family R0942 also exhibited segregation of the variant T allele. In this family, all seven affected individuals were heterozygous for the variant T allele. Out of the 10 unaffected individuals in this family, 6 are wild-type and 4 carry one copy of the variant

T allele. The four unaffected individuals with the variant T allele are 33, 51, 52 and 78 years of age. As in family R1136, no one in family R0942 with the homozygous wild-type G allele had PF.

To determine if segregation of the risk allele occurred by chance in family R0942, Simplified rapid segregation analysis (SISA) was conducted (Møller et al., 2011). This analysis is done by calculating the number of informative meioses in a family, assuming that the variant in question (in this case, the rs35705950 *MUC5B* promoter SNP) is only introduced once into the pedigree (Møller et al., 2011). After obtaining this information, the probability for cosegregation of the phenotype with the genetic variant can be calculated. From performing this analysis, the probability by chance that co-segregation of the variant T allele with pulmonary fibrosis occurred in family R0942 was 1.56%. This confirms it is unlikely that segregation of the *MUC5B* rs35705950 variant T allele with FPF occurred by chance in family R0942. A manuscript with these findings have been submitted on March 2013 to a peer-reviewed journal, *Thorax*, for publication (Appendix I).

**Table 3.4: Clinical variables and findings from rs35705950 *MUC5B* promoter SNP genotyping case vs. control study.**

Clinical variables	All cases (n=110)	Cases: probands only (n=80)	Controls (n=277)
Age (year)*	62.1 ± 12.3	63.1 ± 9.6	61.2 ± 9.4
Gender	M = 67 (60.9%) F = 43 (39.1%)	M = 51 (63.8%) F = 29 (36.2%)	M = 158 (58.5%) F = 112 (41.5%)
Smoking status**	smoker = 17 (15.5%) ex-smoker = 73 (66.3%) occasional = 3 (2.7%) never = 17 (15.5%)	smoker = 6 (7.5%) ex-smoker = 62 (77.5%) occasional = 0 (0.0%) never = 12 (15.0%)	smoker = 33 (12.2%) ex-smoker = 128 (47.4%) occasional = 2 (0.01%) never = 107 (39.6%)
Smoking duration (years)	25.0 ± 16.5	25.0 ± 16.0	26.1 ± 15.6
Pulmonary function tests: mean % of predicted value***			
FVC	84.1% ± 18.8%	83.2% ± 18.1%	N/A
FEV1	83.0% ± 15.9%	82.1% ± 15.8%	N/A
TLC	81.1% ± 16.6%	81.4% ± 15.8%	N/A
DLCO	59.4% ± 17.5%	56.6% ± 14.8%	N/A
Diagnosis by lung biopsy	45 (40.9%)	35 (43.8%)	N/A
Allele frequency:			
Wild type allele (G) frequency	63.60%	62.50%	87.40%
Minor allele (T) frequency	36.40%	37.50%	12.60%
Genotype:			
G/G	39 (35.4%)	28 (35.0%)	211 (76.2%)
G/T	62 (56.4%)	44 (55.0%)	62 (22.4%)
T/T	9 (8.2%)	8 (10.0%)	4 (1.4%)
Odds Ratio (95% CI)			
G/T vs. G/G	5.4 (3.0 - 9.6), P < .001	5.3 (3.1 - 9.2), P < .001	N/A
T/T vs. G/G	12.2 (3.3 - 44.7), P < .001	15.2 (4.3 - 52.6), P < .001	N/A
Chi-squared, P value for Hardy-Weinberg equilibrium	$\chi^2 = 5.22$ , P = .02	$\chi^2 = 2.40$ , P = .12	$\chi^2 = .05$ , P = .82

---

Control population age, gender, smoking status and duration is based on n=270 since questionnaires were not returned by 7 participants. Control population allele frequency, genotypes and odds ratio based on n=277.

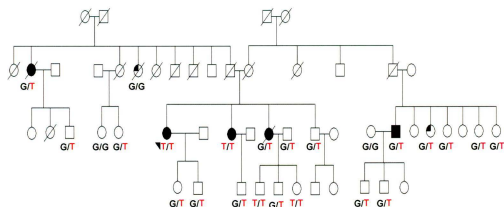
\*For cases, age of diagnosis is used. For controls, age of sample collection is used.

\*\*For smoking status: Ex-smoker (all cases): smoking duration is  $28.9 \pm 13.8$  years with 71.2% having smoked over 20 years. Ex-smoker (probands only): smoking duration is  $28.7 \pm 13.1$  years with 71.0% having smoked over 20 years. Ex-smoker(controls): smoking duration is  $22.8 \pm 14.8$  years with 53.1% having smoked over 20 years. 5 total occasional smokers: only smoke socially with 1/5 having smoked over 20 years.

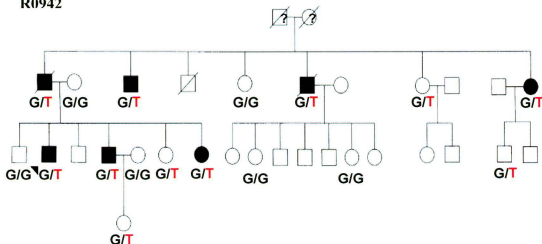
\*\*\*abnormal if FVC, FEV1, TLC or DLCO = <80% predicted

FVC = forced vital capacity, FEV1 = forced expiratory volume exhaled in the 1st second, TLC = total lung capacity, DLCO = diffusing capacity of the lung for carbon monoxide.

### R1136



### R0942



**Figure 3.15: Segregation analysis of *MUC5B* promoter SNP rs35705950 in two FPF families: R1136 and R0942.**

Genotypes for all individuals with DNA in R0942 and R1136 are written on the pedigree below the individual's symbol. Individual with no labels did not have DNA available for genotyping. The variant allele T is highlighted in red. In Family R1136, the two partially shaded individuals represent individuals that are possibly affected but have not been confirmed clinically.



## **4. Discussion**

### **4.1 Summary of the candidate gene analysis**

Previous research carried out by Ms. Laura Edwards was done to determine the genetic loci linked with PF by using a genome-wide scan with microsatellite markers (2006). Fine-mapping was also done using a SNP genome-wide scan to narrow down the ROI linked with FPF in six Newfoundland families (Kamel, 2010). From the ROIs found on chromosome 6p and 16p, eleven genes were sequenced by Kamel (2010). Sequencing of candidate genes from these two ROIs was unsuccessful in finding a variant causing FPF in six Newfoundland families.

The purpose of the candidate gene analysis in this thesis was to identify a relatively highly penetrant genetic variant that causes FPF in Family R0942. From the total of nine candidate genes that were analyzed in this thesis, 28 genetic variants were found as shown in Table 3.2 and Table 3.3. Using a combination of filtering criteria for each genetic variant found, all 28 of these variants were excluded as FPF-causing in Family R0942. The variants were categorized based upon whether they segregated with the disease in family R0942; the Caucasian population frequency for the variant; previous literature reporting the variant and its effects; an bioinformatic analyses on the effects of the variant. All 28 variants found did not segregate with FPF in family R0942 and were previously reported in the dbSNP database. Of these, 26 SNPs were reported to have minor allele frequencies greater than 1% in a Caucasian population sample. All variants found to have a minor allele frequency between 1-5% were re-examined. There was only one variant in this range (*MSLN* c.1230+16G>T, MAF 1000 Genomes: 1.8%). This

variant was excluded as being FPF-causing since it did not segregate with FPF in family R0942. Finally, the two SNPs that did not have a minor allele frequency reported in the literature were not shown to segregate with the disease in Family R0942 and thus also excluded.

Out of all 28 genetic variants found, five were predicted to cause changes in amino acid. These five variants caused a missense change. These variants were filtered again using bioinformatics analysis, population frequency data and literature reviews. During this step, Panther, PolyPhen and SIFT analysis was conducted to see the pathogenicity of the variant. Furthermore, population frequency data was gathered from 1000 Genomes and/or HapMap when available. Finally, a PubMed literature search was conducted to see if additional information on the variant has been published recently. After these additional filtration strategies, these variants were also excluded as FPF-causing in this family.

Overall, there were no variants found via candidate gene analysis that were deemed FPF-causing. This is not altogether unexpected since there are over 300 genes in the ROIs on chromosome 16 and 6 that were not selected for sequencing in this study based on functional/expression data (UCSC Genome Browser, build GRCh37/hg19; Kent et al., 2002). It is plausible that the pathogenic variant is contained in one of these unselected genes or in another gene in a region not highlighted by linkage analysis (see Limitations, section 4.3).

## 4.2 The role of the *MUC5B* promoter SNP in IPF

The purpose of the *MUC5B* rs35705950 SNP genotyping analysis was to test the significance of a recently discovered *MUC5B* promoter variant and its association with risk of developing IPF. From the results of the *MUC5B* promoter SNP genotyping case vs. control study, it is evident that the minor allele, T, of rs35705950 is associated with IPF. This association is seen when both ‘probands only’ and ‘all cases’ are compared to controls (Table 3.4). An additive genotypic effect for the variant T allele with risk of developing PF is also demonstrated by both case versus control analyses. Furthermore, the minor allele segregated with PF in two large affected families. Recent research provides evidence for significantly higher expression of *MUC5B* in patients with IPF compared to controls, *MUC5B* protein presentation in lung tissue of IPF patients, increased *MUC5B* expression for unaffected subjects carrying the minor allele compared to unaffected subjects carrying the wild-type (Seibold et al., 2011). Smoking, which is well-known to be associated with numerous ILDs, appeared to have little effect on the association between all subjects and the minor allele of rs35705950. Coupled with previous research, our findings demonstrate the risk of IPF development with the rs35705950 promoter polymorphism is considerable.

Since the minor allele frequency of rs35705950 is 0.051 (1000 Genomes project, dbSNP), it is not a rare variant in the normal population. It appears that an increased expression of *MUC5B* in the lung, due to the variant allele, increases the risk for IPF development. Furthermore, our results have demonstrated that the risk is an additive genotypic effect (i.e. if someone possesses two copies of the variant T allele, this individual has an increased risk of developing IPF compared to someone possessing one

copy of the variant T allele). Thus, in both familial and sporadic patients, the *MUC5B* promoter polymorphism confers risk to all individuals who carry at least one copy of the minor T allele. When combined with environmental factors such as smoking, occupational hazards, textile dust, and asbestos, it is likely that these individuals will develop IPF in comparison to someone without the rs35705950 minor T allele.

Mucins are a protein family that has a primary ability to form gel surfaces. Mucins are highly glycosylated and are present mainly on the surface of epithelial cells (Devine and McKenzie, 1992). Mucins have a variety of documented functions in both normal cells and tumour cells (Table 4.1). The exact mechanism by which the *MUC5B* promoter polymorphism promotes the development of IPF is not fully understood but a few theories have been postulated.

The main theory is based on IPF development as a result of numerous injuries to the lung and aberrant wound repair processes. From this, it is plausible that significantly increased amounts of *MUC5B* due to the variant allele can impair the defense mechanisms of the mucosal host (Seibold et al., 2011). Impaired defense of the mucosal host would cause excessive injury to the lung since clearance of various particles is reduced, ultimately leading to IPF. Two other prevalent theories involve 1) excess *MUC5B* in respiratory bronchioles interfering with alveolar repair and 2) disruption and/or introduction of transcription-factor binding sites in the *MUC5B* promoter region due to the presence of the variant T allele leading to ectopic production of *MUC5B* protein in a variety of cells in the lung, causing injury to the bronchoalveolar unit (Seibold et al., 2011).

Despite the association of the *MUC5B* promoter polymorphism with PF in the NL cohort, it is possible that the SNP is not associated with the disease. For instance, only 2 out of 12 families with familial pulmonary fibrosis demonstrated segregation of the variant T allele with the disease. It is possible that there are alternative genetic causes for FPF in the 10 families that did not show segregation of *MUC5B* with FPF. Since known genetic causes of FPF (variants in *TERT*, *TERC*, *ABCA3*, *SFTPC*, *SFTPCA2*) have already been tested in these families, there could be a novel genetic variation that is causing FPF in these families. It is also possible that the *MUC5B* rs35705950 SNP is not the sole cause of PF in the familial form and instead serves as a modifier allele. As a modifier allele, the presence of the variant T allele will increase the risk of an individual developing PF but this outcome is not absolute and will differ depending on other genetic variants and environmental factors.

Also, in the two families (R1136 and R0942) that exhibited segregation of the *MUC5B* variant T allele with PF, there are 14 and 4 individuals respectively that carry the variant T allele and are unaffected at last follow up (Figure 3.15). Although this is consistent with PF being a late-onset disease with reduced penetrance, it is also possible that the *MUC5B* rs35705950 variant T allele does not have as large of an effect on the disease as originally demonstrated by Seibold et al (2011).

Taking these alternative explanations into consideration and despite the significant association of the rs35705950 *MUC5B* promoter polymorphism with IPF and the increased risk associated with it, further research must be done to elucidate the exact mechanism of action exerted by this variant allele (see section 4.4). These findings can play a role in identifying individuals who may be at risk to develop PF. Preventative

education could be provided for mutation-positive individuals in hope to lower the chances of developing PF. With the results of this project, preventative measures such as education could be provided at an early age for individuals who carry the *MUC5B* promoter SNP. Changes in lifestyle (i.e. smoking habits, occupational health and hazards) may lower the risk of the mutation-positive individual developing IPF.

#### **4.3 Limitations of this study**

There were a number of limitations in this study. One of the biggest limitations stems from the sequencing methodology used in this study. The seven individuals screened in Family R0942 were investigated only for exonic genetic variants. Functional exons only account for 2% of the mammalian genome (Makalowski, 2000). Therefore, since the majority of the genome is non-exonic, large expanses of sequence was not analyzed and disease-causing variations in these regions of the genome could have been missed.

Besides point mutations and small insertions/deletions in non-exonic parts of the genome, there are a number of other genetic variations that cannot be identified by methods used in this study. For instance, large insertions and/or deletions cannot be identified via Sanger sequencing in a disease that has an autosomal dominant mode of inheritance. Comparative genomic hybridization (CGH) is a method that could be utilized to find any regions that may have been deleted or duplicated. Hence if the FPF phenotype is displayed due to a potential copy number variation in DNA, CGH would be an effective method to determine this. Furthermore, whole genome sequencing (WGS) would

be another effective method to find point mutations, insertions/deletions of all sizes and copy number variations since the data can be analyzed by read depth.

Genetic variations in various promoter or repressor sequences are also not detected by methods used in this study. Most promoter or repressor sequences are located in the 5'UTR, 3'UTR and/or intronic regions. Since the sequence coverage in this study encompasses the exons and <25 base pairs of intron-exon boundaries, promoter or repressor sequences are not covered. There could be genetic variants in these regions that may have consequences on gene expression, transcription and/or regulation. To identify potential variants in these regions, sequencing of the promoter/repressor regions coupled with quantitative gene expression analyses must be carried out. Finally, large genetic changes such as translocations and inversions cannot be detected by methods used in this study.

Another limitation in this study is the lack of functional testing for all genetic variants found. If a variant was found in a candidate gene, criteria for filtering the variant depended on bioinformatics tools such as PolyPhen and SIFT; information from databases such as Ensembl, dbSNP and UCSC Genome Browser; and previous functional testing/expression analyses documented in the literature. In-house functional tests that would aid in determining the consequence of a genetic variant in comparison to wild-type, such as gene expression analyses and protein/enzymatic activity assays, were not performed in this study. Also, for synonymous variants, splice predictor tools were not utilized.

The use of a candidate gene approach is another limitation in this study. To date there are over 15000 genes in the human body. Of these genes, hundreds of these have

functions/roles that are still unknown. Once a region(s) of interest is determined, a candidate gene approach is limited to selecting genes with known functionality. All genes with unknown functions or hypothetical genes were excluded as candidates. It is possible that one of these excluded genes that could harbour a potential FPF-causing variant. Sequence capture of the ROIs would be an ideal solution but this was not performed due to limited resources.

Though it is rare, possible phenocopies may be another limitation in this study. If someone in the R0942 family had PF due to a non-genetic cause, this individual may not have the FPF-causing variant. Since it is assumed that FPF in Family R0942 is due to a relatively strong genetic predisposition, a phenocopy would lead to erroneous conclusions with regards to linkage analysis, fine-mapping and haplotyping.

Additionally, environmental factors such as smoking cigarettes is a confounding factor that could account for higher rates of PF, and this was not controlled for in the study. Finally, using controls from a colorectal cancer project is a limitation in the case-control study presented here since there is no clinical data pertaining to lung disease available for these controls.

#### **4.4 Future work**

Since there are some limitations in the methods of this study that result in potential genetic variants to be missed, whole genome sequencing and CGH should be conducted on a few affected individuals. This would help in identifying all SNPs, insertions, deletions and copy number variations in the entire genome. If the genetic variations found by these methods are properly filtered, the FPF-causing variant could be



identified. Furthermore, these results could be tested in other Newfoundland families and patients that have PF to determine if the same genetic cause accounts for PF in multiple cases. Once the genetic cause is identified, clinical screening for the pathogenic variant can be conducted for at-risk family members. For those that carry the variant, genetic counselors could provide preventative education and support. Also, antifibrotic treatments can be provided by physicians and the mutation-positive individuals can be placed on a lung transplant list earlier to minimize waitlist time and while they are relatively healthy.

Future studies should also carry out various functional tests to elucidate the exact mechanism of action of any pathogenic variants found, such as the *MUC5B* promoter polymorphism. For instance, functional studies such as gene knock-out and/or mutagenesis can be conducted to compare *MUC5B* expression in affected individuals who are mutation carriers of one of the other known IPF-genes in comparison to *MUC5B* expression in affected individuals who have no known mutations in the known-IPF genes. Further studies involving a genetically modified *MUC5B* mouse model can be performed as well. These functional tests are of utmost importance since these tests can also illustrate the pathways and complex interactions involved in the pathogenesis of the disease. Also, there are many non-synonymous/ missense variants found in humans that do not result in the display of any abnormal phenotype- functional studies would clarify the precise role of the variants. Overall, studies that focus on investigating the genetic causes of PF hold great potential for finding results that can be transferred to clinics. It is hopeful that these results could positively impact the diagnosis, prevention, treatments and prognosis of PF.

## **5. Conclusion**

In the candidate gene study, the genetic cause of FPF in Family R0942 was not determined. The candidate gene approach utilizing previous work done (Edwards, 2006; Kamel, 2010) failed to detect a FPF-causing variant in a portion of the ROIs determined by linkage analysis. More candidate gene sequencing and/or next generation sequencing techniques should be carried out to discover this pathogenic variant. However, the variant allele of a *MUC5B* promoter SNP that is associated with developing PF segregated with all affected individuals in Family R0942. This *MUC5B* variant may account for PF in this family by playing a role in the development of the disease. Currently, exome sequencing is also being performed on three individuals in Family R0942 (R0942.0002, R0942.0005, R0942.1002; shown in Figure 3.2). The results from the exome sequencing project will help determine if there is an additional explanation for FPF in this family.

In the case vs. control study, a significant association was observed between the rs35705950 *MUC5B* promoter SNP and development of IPF. Since this association was originally found in a few Caucasian populations in the USA, our study validates this association in the Newfoundland population as well. This significant association observed in multiple independent populations suggests that the rs35705950 *MUC5B* promoter SNP plays an important role in the pathogenesis of IPF. Further studies must be carried out to understand the precise mechanisms of this promoter SNP.

## **References**

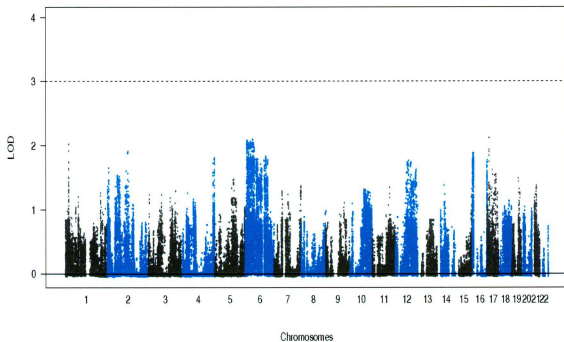
- Alder, J. K., J. J. Chen, et al. (2008). "Short telomeres are a risk factor for idiopathic pulmonary fibrosis." Proc Natl Acad Sci U S A **105**(35): 13051-6.
- Altshuler, D.M., E.S. Lander, et al. (2010). "A Map of Human Genome Variation from Population Scale Sequencing." Nature **467**(7319): 1061-73.
- Antoniou, K. M., G. Soufla, et al. (2010). "Expression analysis of angiogenic growth factors and biological axis CXCL12/CXCR4 axis in idiopathic pulmonary fibrosis." Connect Tissue Res **51**(1): 71-80.
- Armanios, M. Y., J. J. Chen, et al. (2007). "Telomerase mutations in families with idiopathic pulmonary fibrosis." N Engl J Med **356**(13): 1317-26.
- ATS (2000). "American Thoracic Society. Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS)." Am J Respir Crit Care Med **161**(2 Pt 1): 646-64.
- Bantsimba-Malanda, C., J. Marchal-Somme, et al. (2010). "A role for dendritic cells in bleomycin-induced pulmonary fibrosis in mice?" Am J Respir Crit Care Med **182**(3): 385-95.
- Bear, J. C., T. F. Nemece, et al. (1987). "Persistent genetic isolation in outport Newfoundland." Am J Med Genet **27**(4): 807-30.
- Bessler, M., D. B. Wilson, et al. (2010). "Dyskeratosis congenita." FEBS Lett **584**(17): 3831-8.
- Bradley, B., H. M. Branley, et al. (2008). "Interstitial lung disease guideline: the British Thoracic Society in collaboration with the Thoracic Society of Australia and New Zealand and the Irish Thoracic Society." Thorax **63 Suppl 5**: v1-58.
- Bullard, J. E., S. E. Wert, et al. (2005). "ABCA3 mutations associated with pediatric interstitial lung disease." Am J Respir Crit Care Med **172**(8): 1026-31.
- Chibbar, R., F. Shih, et al. (2004). "Nonspecific interstitial pneumonia and usual interstitial pneumonia with mutation in surfactant protein C in familial pulmonary fibrosis." Mod Pathol **17**(8): 973-80.
- Cosgrove, G. P., K. K. Brown, et al. (2004). "Pigment epithelium-derived factor in idiopathic pulmonary fibrosis: a role in aberrant angiogenesis." Am J Respir Crit Care Med **170**(3): 242-51.
- Crossno, P., Polosukhin, V, et al. (2010). "Identification of Early Interstitial Lung Disease in an Individual With Genetic Variations in ABCA3 and SFTPC." Chest **137**(4): 969-73.
- Davies, H. T. (1998). "Interpreting measures of treatment effect." Hosp Med **59**(6): 499-501.
- Devine, P. L. and I. F. McKenzie (1992). "Mucins: structure, function, and associations with malignancy." Bioessays **14**(9): 619-25.
- Diaz de Leon, A., J. T. Cronkhite, et al. (2010). "Telomere lengths, pulmonary fibrosis and telomerase (TERT) mutations." PLoS One **5**(5): e10680.
- du Bois, R. M. (2006). "Genetic factors in pulmonary fibrotic disorders." Semin Respir Crit Care Med **27**(6): 581-8.

- Edwards, L. (2006). "The search for genetic loci linked to pulmonary fibrosis in Newfoundland." Honours Thesis, Memorial University.
- Erlich, H. A. (1989). "Polymerase chain reaction." *J Clin Immunol* **9**(6): 437-47.
- Fertig, N., R. T. Domsic, et al. (2009). "Anti-U11/U12 RNP antibodies in systemic sclerosis: a new serologic marker associated with pulmonary fibrosis." *Arthritis Rheum* **61**(7): 958-65.
- Greenberg, D. A., P. Abreu, et al. (1998). "The power to detect linkage in complex disease by means of simple LOD-score analyses." *Am J Hum Genet* **63**(3): 870-9.
- Greene, K. E., T. E. King, Jr., et al. (2002). "Serum surfactant proteins-A and -D as biomarkers in idiopathic pulmonary fibrosis." *Eur Respir J* **19**(3): 439-46.
- Gribbin, J., R. B. Hubbard, et al. (2006). "Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK." *Thorax* **61**(11): 980-5.
- Gross, T. J. and G. W. Hunninghake (2001). "Idiopathic pulmonary fibrosis." *N Engl J Med* **345**(7): 517-25.
- Henke, M. O., A. Renner, et al. (2004). "MUC5AC and MUC5B Mucins Are Decreased in Cystic Fibrosis Airway Secretions." *Am J Respir Cell Mol Biol* **31**(1): 86-91.
- Horne, B. D., A. Malhotra, et al. (2003). "Comparison of linkage analysis methods for genome-wide scanning of extended pedigrees, with application to the TG/HDL-C ratio in the Framingham Heart Study." *BMC Genet* **4 Suppl 1**: S93.
- Kawaguchi, M., M. Adachi, et al. (2004). "IL-17 cytokine family." *J Allergy Clin Immunol* **114**(6): 1265-73; quiz 1274.
- Kamel, F. (2010). "Determining the Genetic Etiology of Familial Pulmonary Fibrosis in Six Newfoundland Families" Masters thesis, Memorial University.
- Kent, W. J., Sugnet, C. W., et al. (2002). "The human genome browser at UCSC." *Genome Res* **12**(6): 996-1006.
- King, T. E., Jr., A. Pardo, et al. (2011). "Idiopathic pulmonary fibrosis." *Lancet* **378**(9807): 1949-61.
- Kirkham, S., U. Kolsum, et al. (2008). "MUC5B is the major mucin in the gel phase of sputum in chronic obstructive pulmonary disease." *Am J Respir Crit Care Med* **178**(10): 1033-9.
- Lathrop, G. M., J. M. Lalouel, et al. (1984). "Strategies for multilocus linkage analysis in humans." *Proc Natl Acad Sci U S A* **81**(11): 3443-6.
- Li, W., Q. Zhang, et al. (2003). "Hermansky-Pudlak syndrome type 7 (HPS-7) results from mutant dysbindin, a member of the biogenesis of lysosome-related organelles complex 1 (BLOC-1)." *Nat Genet* **35**(1): 84-9.
- Makalowski, W. (2000). "Genomic scrap yard: how genomes utilize all that junk." *Gene* **259**(1-2): 61-7.
- Martinez, P. and M. A. Blasco (2011). "Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins." *Nat Rev Cancer* **11**(3): 161-76.
- McCormack, F. X., T. E. King, Jr., et al. (1991). "Idiopathic pulmonary fibrosis. Abnormalities in the bronchoalveolar lavage content of surfactant protein A." *Am Rev Respir Dis* **144**(1): 160-6.
- Merner, N. D., K. A. Hodgkinson, et al. (2008). "Arrhythmogenic right ventricular cardiomyopathy type 5 is a fully penetrant, lethal arrhythmic disorder caused by a missense mutation in the TMEM43 gene." *Am J Hum Genet* **82**(4): 809-21.

- Nakashima, T., A. Yokoyama, et al. (2008). "Suppressor of cytokine signaling 1 inhibits pulmonary inflammation and fibrosis." *J Allergy Clin Immunol* **121**(5): 1269-76.
- Nogee, L. M., A. E. Dunbar, 3rd, et al. (2001). "A mutation in the surfactant protein C gene associated with familial interstitial lung disease." *N Engl J Med* **344**(8): 573-9.
- Ono, S., T. Tanaka, et al. (2011). "Surfactant protein C G100S mutation causes familial pulmonary fibrosis in Japanese kindred." *Eur Respir J* **38**(4): 861-9.
- Park, J. H., D. K. Kim, et al. (2007). "Mortality and risk factors for surgical lung biopsy in patients with idiopathic interstitial pneumonia." *Eur J Cardiothorac Surg* **31**(6): 1115-9.
- Pierson, D. M., D. Ionescu, et al. (2006). "Pulmonary fibrosis in hermansky-pudlak syndrome. a case report and review." *Respiration* **73**(3): 382-95.
- Raghu, G., D. Weycker, et al. (2006). "Incidence and prevalence of idiopathic pulmonary fibrosis." *Am J Respir Crit Care Med* **174**(7): 810-6.
- Rahman, P., A. Jones, et al. (2003). "The Newfoundland population: a unique resource for genetic investigation of complex diseases." *Hum Mol Genet* **12 Spec No 2**: R167-72.
- Ramensky, V., P. Bork, et al. (2002). "Human non-synonymous SNPs: server and survey." *Nucleic Acids Res* **30**(17): 3894-900.
- Rodriguez, S., T. R. Gaunt, et al. (2009). "Hardy-Weinberg equilibrium testing of biological ascertainment for Mendelian randomization studies." *Am J Epidemiol* **169**(4): 505-14.
- Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." *Methods Mol Biol* **132**: 365-86.
- Seibold, M. A., A. L. Wise, et al. (2011). "A common MUC5B promoter polymorphism and pulmonary fibrosis." *N Engl J Med* **364**(16): 1503-12.
- Selman, M., T. E. King, et al. (2001). "Idiopathic pulmonary fibrosis: prevailing and evolving hypotheses about its pathogenesis and implications for therapy." *Ann Intern Med* **134**(2): 136-51.
- Steele, M. P. and K. K. Brown (2007). "Genetic predisposition to respiratory diseases: infiltrative lung diseases." *Respiration* **74**(6): 601-8.
- Strachan, T. and A. P. Read (1999). "Human Molecular Genetics 2."
- Stuckless, S., P. S. Parfrey, et al. (2007). "The phenotypic expression of three MSH2 mutations in large Newfoundland families with Lynch syndrome." *Fam Cancer* **6**(1): 1-12.
- Swigris, J. J. and K. K. Brown (2010). "The role of endothelin-1 in the pathogenesis of idiopathic pulmonary fibrosis." *BioDrugs* **24**(1): 49-54.
- Takahashi, H., T. Fujishima, et al. (2000). "Serum surfactant proteins A and D as prognostic factors in idiopathic pulmonary fibrosis and their relationship to disease extent." *Am J Respir Crit Care Med* **162**(3 Pt 1): 1109-14.
- Takai, H., R. C. Wang, et al. (2007). "Tel2 regulates the stability of PI3K-related protein kinases." *Cell* **131**(7): 1248-59.
- Tan, K., K. Kajino, et al. (2010). "Mesothelin (MSLN) promoter is hypomethylated in malignant mesothelioma, but its expression is not associated with methylation status of the promoter." *Hum Pathol* **41**(9): 1330-8.

- Tsakiri, K. D., J. T. Cronkhite, et al. (2007). "Adult-onset pulmonary fibrosis caused by mutations in telomerase." Proc Natl Acad Sci U S A **104**(18): 7552-7.
- van Moersel, C. H., M. F. van Oosterhout, et al. (2010). "Surfactant protein C mutations are the basis of a significant portion of adult familial pulmonary fibrosis in a dutch cohort." Am J Respir Crit Care Med **182**(11): 1419-25.
- Vece, T. J., M. G. Schecter, et al. (2012). "Rapid and Progressive Pulmonary Fibrosis in 2 Families with DNA Repair Deficiencies of Undetermined Etiology." J Pediatr.
- Waghray, M., Z. Cui, et al. (2005). "Hydrogen peroxide is a diffusible paracrine signal for the induction of epithelial cell death by activated myofibroblasts." FASEB J **19**(7): 854-6.
- Wang, P. P., E. Dicks, et al. (2009). "Validity of random-digit-dialing in recruiting controls in a case-control study." Am J Health Behav **33**(5): 513-20.
- Wang, Y., P. J. Kuan, et al. (2009). "Genetic defects in surfactant protein A2 are associated with pulmonary fibrosis and lung cancer." Am J Hum Genet **84**(1): 52-9.
- Wells, A. U., S. R. Desai, et al. (2003). "Idiopathic pulmonary fibrosis: a composite physiologic index derived from disease extent observed by computed tomography." Am J Respir Crit Care Med **167**(7): 962-9.
- Woods, M. O., H. B. Younghusband, et al. (2010). "The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease." Gut **59**(10): 1369-77.
- Wu, C., C. Orozco, et al. (2009). "BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources." Genome Biol **10**(11): R130.
- Xie, Y. G., H. Zheng, et al. (2002). "A founder factor VIII mutation, valine 2016 to alanine, in a population with an extraordinarily high prevalence of mild hemophilia A." Thromb Haemost **87**(1): 178-9.
- Yamano, G., H. Funahashi, et al. (2001). "ABCA3 is a lamellar body membrane protein in human lung alveolar type II cells." FEBS Lett **508**(2): 221-5.
- Young, L. R., L. M. Nogee, et al. (2008). "Usual interstitial pneumonia in an adolescent with ABCA3 mutations." Chest **134**(1): 192-5.
- Young, T. L., L. Penney, et al. (1999). "A fifth locus for Bardet-Biedl syndrome maps to chromosome 2q31." Am J Hum Genet **64**(3): 900-4.
- Zhang, G. Y., T. Liao, et al. (2011). "MUC5B promoter polymorphism and pulmonary fibrosis." N Engl J Med **365**(2): 178; author reply 178-9.

**Appendix A: 2-point parametric linkage LOD scores on 500k SNPs for R0942,  
dominant model**



This is a Manhattan plot received by Mr. Fady Kamel. This data and the accompanying SNP genome-wide analysis for R0942 were re-visited by me in order to confirm and fine-map the ROIs.

## Appendix B: Candidate gene tables for ROIs on chromosome 6p and 16p

Gene	Position	Function	Expression	Candidate ?
HULC	chr6:8597441-8599080	highly up-regulated in liver cancer (non-protein coding RNA): a novel gene with striking up-regulation in hepatocellular carcinoma	high: liver	very unlikely
TFAP2A	chr6:10504902-10520593	transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha): defects in this gene are a cause of branchiooculofacial syndrome (BOFS)	high: retina, placenta	sequenced (Kamel)
GCNT2	chr6:10629554-10737587	glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (I blood group)--> encodes enzyme responsible for formation of blood group I antigen, Mutations:have been associated with adult i blood group phenotype	ubiquitous	unlikely
PAK1IP1	chr6:10803174-10817956	PAK1 interacting protein 1; Mutation analysis shows PIP1 interacts specifically with the N-terminal regulatory domain of PAK1, which contains NCK-binding motif, a dimerization domain, an inhibitory switch domain and a kinase inhibitor segment	high: prostate	unlikely
TMEM14C	chr6:10831134-10839348	Transmembrane protein 14C; putative mitochondrial transporter	n/a	unlikely
TMEM14B	chr6:10855981-10865200	Transmembrane protein 14B, unknown	n/a	unlikely
MAK	chr6:10870942-10939096	male germ cell-associated kinase: product of this gene is a serine/threonine protein kinase related to kinases involved in cell cycle regulation	Almost exclusively in testis, primarily in germ cells	very unlikely
GCM2	chr6:10981442-10990084	glial cells missing homolog 2 (Drosophila): thought to act as a binary switch between neuronal and glial cell determination, Mutations in this gene are associated with hypoparathyroidism	high: skeletal muscle, liver, heart... low: ovary	very unlikely
SYCP2L	chr6:10995050-11082527	synaptonemal complex protein 2-like: <i>loci</i> associated with age at menarche and age at natural menopause (study in Nature Genetics)	unknown	unlikely
ELOVL2	chr6:11088979-11152610	ELONGATION OF VERY LONG CHAIN FATTY ACIDS-LIKE 2: when overexpressed in preadipocyte cell lines: increased TAG synth, enhanced bulidup of lipid droplets, induced expression of diacylglycerol acyltransferase-2 etc	high: fetal liver, liver, heart	very unlikely
LOC221710	chr6:11202252-11246955	hypothetical protein LOC221710, function unknown	unknown	unlikely



<b>ERVFRDE1</b>	chr6:11210708-11220057	most HERVs (endogenous RVs) are nonfunctional, the HERV-W and HERV-FRD envelope (env) proteins can induce cell-cell fusion when expressed in cells possessing appropriate receptors	high: placenta	unlikely
<b>NEDD9</b>	chr6:11291517-11340901	neural precursor cell expressed, developmentally down-regulated 9; may be an important linking element between extracellular signaling and regulation of the cytoskeleton, linked in melanomas....	high: in kidney, lung and placenta... lung: highest expression, >>10xM	possible
<b>TMEM170B</b>	chr6:11646497-11691743	Transmembrane protein 170B, the TMEM170B gene is conserved in chimpanzee, dog, mouse, rat, chicken, and zebrafish..	specific functions and expression pattern unknown (BioGPS)	unlikely
<b>HIVEP1</b>	chr6:12120710-12273218	the proteins bind specific DNA sequences, including kappa-B motif in the promoters/enhancer regions of several genes and viruses (including HIV)..... has some effects in pulmonary hypertension....	high: cardiac myocytes, liver, heart... lung: median	possible but unlikely
<b>EDN1</b>	chr6:12398515-12405413	This peptide is a potent vasoconstrictor and is produced by vascular endothelial cells. Has been implicated in wrt IPF.	High: lung	sequenced
<b>PHACTR1</b>	chr6:12825819-13395507	phosphatase and actin regulator 1..... recombinant rat Phactr1 inhibited PP1 enzymatic activity in a concentration-dependent manner	high: placenta, obo... lung: median	unlikely
<b>TBC1D7</b>	chr6:13413163-13436766	a role in regulating cell growth and differentiation, Activation of an oncogenic TBC1D7 (TBC1 domain family, member 7) protein in pulmonary carcinogenesis	high: heart... low. kidney, liver, placenta	possible
<b>GFOD1</b>	chr6:13471798-13595766	glucose-fructose oxidoreductase domain containing 1	high: whole brain.... lung: ~1.5x median	very unlikely
<b>SIRT5</b>	chr6:13682771-13713949	SIRT5 belongs to the sirtuin family of NAD(+)-dependent deacetylases and mono-ADP-ribosyltransferases. Sirtuins control a variety of cellular processes, such as aging, metabolism, and gene silencing	widely expressed in fetal and adult tissues	possible but unlikely
<b>NOL7</b>	chr6:13723538-13729106	NOL7-transfected cells had decreased angiogenesis in a mouse model (main role=angiogenesis)	high: adrenal/thyroid gland, heart, muscle	unlikely
<b>RANBP9</b>	chr6:13729709-13819775	the protein that binds RAN, a small GTP binding prtn (part of RAS superfamily) key for translocation of RNA/pr's through the nuclear pore complex...also interacts with other prtns(met proto-oncogene, homeodomain interacting protein kinase 2, androgen R..	high: testes	very unlikely
<b>CCDC90A</b>	chr6:13898999-	coiled-coil domain containing 90A, mitochondrial precursor?	high:	unlikely

	13922768		lymphoblasts, pineal, CD34 cells	
<b>RNF182</b>	chr6:14032656-14088219	ring finger protein 182: involved in protein ubiquitylation; protein, zinc ion, metal ion binding, ligase activity	ubiquitously expressed	unlikely
<b>CD83</b>	chr6:14225844-14245127	involved in cell-cell adhesion and signal transduction, immune response, regulation of interleukin-4, 10, and 2..."A role for dendritic cells in bleomycin-induced pulmonary fibrosis in mice?"	high: immune system, lymph nodes, spleen, and tonsils... lung: median	sequenced
<b>JARID2</b>	chr6:15354506-15630232	jumonji, AT rich interactive domain 2: overexpression of JMJ appeared to inhibit cell growth, whereas Jmj-deficient mice had cell growth enhancement	high: dorsal root ganglia neurons, cerebral cortex.... lung: median	unlikely
<b>DTNBP1</b>	chr6:15631785-15771268	dystrobrevin binding protein 1: proteins are associated with Hermansky-Pudlak syndrome and schizophrenia, have widespread functions in body-	ubiquitously expressed, high in whole brain	sequenced
<b>LOC401242</b>	chr6:28935381-28939433	hypothetical LOC401242, non-coding RNA:	n/a	very unlikely
<b>TRIM27</b>	chr6:28978758-28999747	The TRIM motif includes three zinc-binding domains, a RING, a B-box type 1/2, and a coiled-coil region.. localizes to nuclear matrix. It interacts with the enhancer of polycomb protein and represses gene transcription. It is also thought to be involved i	higher in CD cells (BioGPS)	unlikely
<b>ZNF311</b>	chr6:29070573-29081016	zinc finger protein 311 (ZNF311), mRNA.	n/a	unlikely
<b>LOC100129636</b>	chr6:29111979-29152496	hypothetical protein LOC100129636 (LOC100129636), mRNA	n/a	unlikely
<b>OR2W1</b>	chr6:29119969-29120931	olfactory receptor, family 2, subfamily W, member 1: The olfactory receptor proteins are members of a large family of G-protein-coupled receptors (GPCR) arising from single coding-exon genes	high: ganglion cells	very unlikely
<b>OR2B3</b>	chr6:29161964-29163069	Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell. The olfactory receptor proteins are members of a large family of G-protein-coupled receptors (GPCR)	n/a	unlikely
<b>OR2J3</b>	chr6:29187647-29188582			
<b>OR2J2</b>	chr6:29249290-29250330			
<b>OR14J1</b>	chr6:29382446-			

OR5V1	29383411 chr6:29430986-29432033			
OR12D3	chr6:29449179-29451047			
OR12D2	chr6:29472395-29473427			
OR11A1	chr6:29501260-29503488			
OR10C1	chr6:29515772-29516710			
OR2H1	chr6:29534209-29540078			
MAS1L	chr6:29562522-29563658	MAS1 oncogene-like: The authors concluded that MRGs may regulate nociceptor function and/or development, including the sensation or modulation of pain	high: specific sensory neurons	unlikely
LOC100507362	chr6:29605162-29609324	hypothetical LOC100507362 (LOC100507362), transcript variant 1, non-coding RNA	n/a	unlikely
UBD	chr6:29631368-29635681	ubiquitin D: encoding a protein similar to diubiquitin. The 1.1-kb transcript was detected only in B-cell lines transformed by Epstein-Barr virus.	high: B lymphoblasts	unlikely
SNORD32B	chr6:29658008-29658084	small nucleolar RNA, C/D box 32B	n/a	unlikely
OR2H2	chr6:29663662-29664724	olfactory receptor, family 2, subfamily H, member 2	high: ganglion cells	unlikely
GABBR1	chr6:29677984-29708941	GABA(B) receptor 1 gene is mapped to chromosome 6p21.3 within the HLA class I region close to the HLA-F gene. Susceptibility loci for multiple sclerosis, epilepsy, and schizophrenia have also been mapped in this region.	high: tonsil, B lymphoblasts	unlikely
MOG	chr6:29732737-29748128	myelin oligodendrocyte glycoprotein: a membrane protein expressed on the oligodendrocyte cell surface and the outermost surface of myelin sheaths. Due to this localization, it is a primary target antigen involved in immune-mediated demyelination.	high: hypothalamus, spinal cord, brain	unlikely
ZFP57	chr6:29748148-29752910	zinc finger protein 57 homolog (mouse): Studies in mouse suggest that this protein may function as a transcriptional repressor. Mutations in this gene have been associated with transient neonatal diabetes mellitus type 1	n/a	unlikely
HLA-F	chr6:29799096-29802282	major histocompatibility complex, class I, F: Unlike most other HLA heavy chains, this molecule is localized in the endoplasmic reticulum and Golgi apparatus, with a small amount present at the cell surface in some cell types. It contains a divergent peptide-binding groove, and is	highest: whole blood, high: lung	unlikely

		thought to bind a restricted subset of peptides for immune presentation. This gene exhibits few variants.		
<b>LOC285830</b>	chr6:29802357-29824805	hypothetical LOC285830, non-coding RNA	high: ovary, lymphnode	unlikely
<b>IFITM4P</b>	chr6:29826563-29826904	interferon induced transmembrane protein 4 pseudogene, non-coding RNA	n/a	unlikely
<b>HCG4</b>	chr6:29866787-29868829	HLA complex group 4 (HCG4), non-coding RNA	ubiquitous	unlikely
<b>HLA-G</b>	chr6:29902735-29906878	major histocompatibility complex, class I, G: belongs to the HLA class I heavy chain paralogues	high: lung, blood, fetal derived placental cells	unlikely
<b>HCP5P10</b>	chr6:29947651-29949539	HLA complex P5 pseudogene 10 (HCP5P10), non-coding RNA	n/a	unlikely
<b>HLA-H</b>	chr6:29963362-29966835	major histocompatibility complex, class I, H, non-coding RNA: represents a transcribed pseudogene, possibly derived from HLA-A. This gene displays extensive variation.	n/a	unlikely
<b>HCG2P7</b>	chr6:29974787-29978410	HLA complex group 2 pseudogene 7 (HCG2P7), non-coding RNA	high: skin, ganglion cells	unlikely
<b>HCG4P6</b>	chr6:30000348-30001407	HLA complex group 4 pseudogene 6 (HCG4P6), non-coding RNA	ubiquitous	unlikely
<b>HLA-A</b>	chr6:30018288-30021640	major histocompatibility complex, class I, A: Class I molecules: central role in immune system by presenting peptides derived from the ER lumen. Polymorphisms within exon 2/3 are responsible for peptide binding specificity of each class I molecule. Typing for these polymorphisms is done for bone marrow/ kidney transplantation.	ubiquitous	unlikely
<b>HCG9</b>	chr6:30050871-30054156	HLA complex group 9 (non-protein coding): its function has not been determined.	high: liver, heart	unlikely
<b>ZNRD1-AS1</b>	chr6:30076767-30136940	ZNRD1 antisense RNA 1 (non-protein coding), non-coding RNA	high: pineal	unlikely
<b>HLA-J</b>	chr6:30081727-30085712	major histocompatibility complex, class I, J: transcribed pseudogene, possibly derived from HLA-A	highest: whole blood, high: lung	unlikely
<b>ZNRD1</b>	chr6:30137015-30140665	zinc ribbon domain containing 1: encodes a protein with similarity to the Saccharomyces cerevisiae Rpa12p subunit of RNA polymerase I.	n/a	unlikely
<b>PPP1R11</b>	chr6:30142911-30146087	protein phosphatase 1, regulatory subunit 11: specific inhibitor of protein phosphatase-1 (PP1) with a differential sensitivity toward the metal-independent and metal-dependent forms of PP1	high: blood, brain, lung, placenta	unlikely
<b>RNF39</b>	chr6:30146022-	ring finger protein 39: Studies of a similar rat protein suggest that this	high: superior	unlikely

	30151607	gene encodes a protein that plays a role in an early phase of synaptic plasticity.	cervical ganglion, trachea	
TRIM31	chr6:30178653-30188846	tripartite motif containing 31: member of the tripartite motif (TRIM) family. The TRIM motif includes three zinc-binding domains, a RING, a B-box type 1 and a B-box type 2, and a coiled-coil region. The protein localizes to both the cytoplasm and the nucleus. Its function has not been identified.	high: skeletal muscle, heart	unlikely
TRIM40	chr6:30212489-30224491	tripartite motif containing 40	n/a	unlikely
TRIM10	chr6:30227703-30236690	tripartite motif containing 10: localizes to cytoplasmic bodies. Studies in mice suggest that this protein plays a role in terminal differentiation of erythroid cells.	high: early erythroid cells	unlikely
TRIM15	chr6:30238962-30248452	tripartite motif containing 15: localizes to the cytoplasm.	n/a	unlikely
TRIM26	chr6:30260211-30289132	tripartite motif containing 26: localizes to cytoplasmic bodies. Function unknown, but RING domain suggests that the protein may have DNA-binding activity.	high: B lymphoblasts	unlikely
HLA-L	chr6:30335318-30342707	major histocompatibility complex, class I, L, pseudogene; non-coding RNA	n/a	unlikely
HCG18	chr6:30363153-30402906	HLA complex group 18; non-coding RNA	n/a	unlikely
TRIM39	chr6:30402987-30419485	tripartite motif containing 39. function of this protein has not been identified.	n/a	unlikely
RPP21	chr6:30420885-30422614	ribonuclease P/MRP 21kDa subunit: a protein subunit of nuclear ribonuclease P, which processes the 5-prime leader sequence of precursor tRNAs	High: heart, B lymphoblasts, T cells	unlikely
HLA-E	chr6:30565162-30569961	major histocompatibility complex, class I, E: belongs to the HLA class I heavy chain paralogues.	high: whole blood, lung	unlikely
GNL1	chr6:30617134-30633350	guanine nucleotide binding protein-like 1: shows a high degree of similarity with its mouse counterpart.	high: pineal	unlikely
PRR3	chr6:30632465-30640452	proline rich 3: function not well known in humans	high: B lymphoblasts	unlikely
ABCF1	chr6:30647149-30667288	ATP-binding cassette, sub-family F (GCN20), member 1: protein found in tumor necrosis factor-alpha-stimulated synoviocytes (cells of alveolar-capillary membrane)	high: B lymphoblasts	possible

Gene	Position	Function	Expression	Candidate ?
LOC100288778	chr16:4043-7324	WAS protein family homolog 1 pseudogene (LOC100288778), non-coding RNA	n/a	unlikely
POLR3K	chr16:36979-43632	polymerase (RNA) III (DNA directed) polypeptide K: the polymerase responsible for synthesizing transfer and small ribosomal RNAs in eukaryotes	high: lymphoblasts, CD cells.... Lung: median expression	unlikely
SNRNP25	chr16:43829-47669	small nuclear ribonucleoprotein 25kDa (U11/U12)--> study: Anti-U11/U12 RNP antibodies are present in the sera of approximately 3% of patients with SSc and are a marker for lung fibrosis, which is often severe.	high: heart, liver, brain	sequenced
RHBDF1	chr16:48058-62629	integral to membrane, serine-type endopeptidase activity,	high: OFB, lung, prostate, placenta	poss but unlikely
MPG	chr16:67018-75843	N-methylpurine-DNA glycosylase: removes a diverse group of damaged bases from DNA, including cytotoxic and mutagenic alkylation adducts of purines	high: thyroid, prostate, lung	unlikely
NPRL3	chr16:75804-128672	nitrogen permease regulator-like 3: Galactose-binding domain, cellular component	high: CD71/early erythroid	unlikely
HBZ HBM HBA2 HBA1 HBQ1	chr16:142854-144504 chr16:155973-156767 chr16:162846-163709 chr16:166679-167520 chr16:170333-171178	he zeta-globin polypeptide is synthesized in the yolk sac of the early embryo, while alpha-globin is produced throughout fetal and adult life. The zeta-globin gene is a member of the human alpha-globin gene cluster that includes five functional genes and two pseudogenes. The order of genes is: 5' - zeta - pseudozeta - mu - pseudoalpha-1 - alpha-2 - alpha-1 - theta1 - 3'	Hb levels decline sometimes in IPF	possible
LUC7L	chr16:178971-219450	LUC7-like ( <i>S. cerevisiae</i> ); may represent a mammalian heterochromatic gene, encoding a putative RNA-binding protein similar to the yeast Luc7p subunit of the U1 snRNP splicing complex that is normally required for 5-prime splice site selection	n/a	unlikely
ITFG3	chr16:224802-256120	integrin alpha FG-GAP repeat containing 3 (ITFG3), mRNA; integral to membrane. FG-GAP repeats are found in N terminus of integrin alpha chains, a region that is important for ligand binding. A putative Ca2+ binding motif is found in some of the repeats	high: thyroid, prostate, retina... lung: ~3xM	unlikely
RGS11	chr16:258311-265915	act as GTPase-activating proteins (GAPs) on the alpha subunits of signal-transducing G proteins. RGS proteins can serve as negative regulators of G protein-mediated signaling pathways by speeding the inactivation of GTP-bound G-alpha subunits.	higher in pineal gland	unlikely
ARHGDI	chr16:270607-273004	Rho GDP dissociation inhibitor (GDI) gamma (ARHGDI): The GDI-dissociation inhibitors (GDIs) play a primary role in modulating the	high: brain region	very unlikely

activation of GTPases by inhibiting the exchange of GDP for GTP				
<b>PDIA2</b>	chr16:273119-277210	protein disulfide isomerase family A, member 2: are endoplasmic reticulum (ER) resident proteins that catalyze protein folding and thiol-disulfide interchange reactions	higher in pancreas	unlikely
<b>AXIN1</b>	chr16:277441-342677	encodes a cytoplasmic protein which contains a regulation of G-protein signaling (RGS) domain and a dishevelled and axin (DIX) domain. Mutations in this gene have been associated with hepatocellular carcinoma, hepatoblastomas, ovarian endometrioid adenocarcinomas, and medullablastomas... <i>reduced exp in human lung carcinomas</i>	n/a	possible but unlikely
<b>MRPL28</b>	chr16:357385-360570	mitochondrial ribosomal protein L28: Mammalian mitochondrial ribosomal proteins are encoded by nuclear genes and help in parietal synthesis within the mitochondria	ubiquitous	unlikely
<b>TMEM8A</b>	chr16:360777-371951	transmembrane protein 8A	high: placenta... lung: 4xM	unlikely
<b>NME4</b>	chr16:387193-390755	non-metastatic cells 4, protein expressed in (NME4), nuc gene encoding mitochon. protein: ubiquitous enzymes that catalyze transfer of gamma-phosphates, via a phosphohistidine intermediate, between nucleoside and dioxynucleoside tri- and diphosphates	high: prostate, heart, liver, small intestine, and skeletal muscle tissues	unlikely
<b>DECR2</b>	chr16:391859-402488	2,4-dienoyl CoA reductase 2, peroxisomal (DECR2)	high: liver	unlikely
<b>RAB11FIP3</b>	chr16:415669-512482	RAB11 family interacting protein 3 (class II): are regulatory roles in the formation, targeting, and fusion of intracellular transport vesicles. RAB11FIP3 is one of many proteins that interact with and regulate Rab GTPases	high: kidney	unlikely
<b>SOLH</b>	chr16:517857-544637	small optic lobes homolog (Drosophila): protein containing zinc-finger-like repeats and a calpain-like protease domain. The encoded protein may function as a transcription factor, RNA-binding protein, or in protein-protein interactions during visual system development	n/a	very unlikely
<b>MIR3176</b>	chr16:533278-533367	microRNA 3176: (miRNAs) are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNAs	n/a	unlikely
<b>NHLRC4</b>	chr16:557033-559496	NHL repeat containing 4:	n/a	unlikely
<b>PIGQ</b>	chr16:560005-574109	phosphatidylinositol glycan anchor biosynthesis, class Q: This gene is involved in the first step in glycosylphosphatidylinositol (GPI)-anchor biosynthesis	high: CD71/early erythroid	unlikely
<b>RAB40C</b>	chr16:579670-619274	member RAS oncogene family (RAB40C):	higher: brain neurons	unlikely
<b>WFIKK1</b>	chr16:621013-624117	WAP, follistatin/kazal, immunoglobulin, kunitz and netrin domain	high: pancreas,	poss but

		containing 1: These domains have been implicated frequently in inhibition of various types of proteases... found in squamous cell lung carcinoma	kidney, liver, placenta, and lung	unlikely
<b>C16orf13</b>	chr16:624430-626348	chromosome 16 open reading frame 13 (C16orf13), transcript variant 7, mRNA	n/a	n/a
<b>WDR90</b>	chr16:639364-657830	WD repeat domain 90	n/a	unlikely
<b>RHOT2</b>	chr16:658134-664172	ras homolog gene family, member T2 (RHOT2), nuclear gene encoding mitochondrial protein... role in mitochon. Distribution and transport	high: retina, pineal, CD cells	unlikely
<b>RHBDL1</b>	chr16:666076-668268	rhomboid, veinlet-like 1 (Drosophila)... implicated in the modulation of epidermal growth factor receptor.	high: brain, kidney, heart	unlikely
<b>STUB1</b>	chr16:670116-672769	ubiquitin ligase/cochaperone that participates in protein quality control by targeting a broad range of chaperone protein substrates for degradation	high: striated muscle(heart, skeletal muscle).... Little in lung	unlikely
<b>JMJD8</b>	chr16:671668-674440	jumonji domain containing 8: Transcription factor jumonji/aspartyl beta-hydroxylase:	high: smooth and cardiac muscle	unlikely
<b>WDR24</b>	chr16:674703-680401	WD repeat domain 24: G-protein beta WD-40 repeat...	high: thyroid	unlikely
<b>FBXL16</b>	chr16:682501-695826	F-box and leucine-rich repeat protein 16: they interact with ubiquitination targets through other protein interaction domains	high: amygdala	very unlikely
<b>METRNL</b>	chr16:705174-707481	Meteorin regulates glial cell differentiation and promotes the formation of axonal networks during neurogenesis	high: whole brain	very unlikely
<b>FAM173A</b>	chr16:711159-712591	family with sequence similarity 173, member A: hypothetical protein LOC65990	high: thyroid, lung, liver, heart, brain	n/a
<b>CCDC78</b>	chr16:712583-716474	coiled-coil domain containing 78:	high: colon, SI	unlikely
<b>HAGHL</b>	chr16:716959-719716	hydroxyacylglutathione hydrolase-like: hydrolase activity, metal ion binding	n/a	unlikely
<b>NARFL</b>	chr16:719770-730998	nuclear prelamin A recognition factor-like: shares similarity with several bacterial iron-only hydrogenases, including conservation of a distinctive active site iron sulfur cluster, termed the H cluster.... Involved in response to hypoxia, o2 homeostasis etc	high: liver, heart, skeletal muscle	unlikely
<b>MSLN</b>	chr16:750766-758866	precursor protein that is cleaved into 2 products, megakaryocyte potentiating factor and mesothelin. Megakaryocyte potentiation factor functions as a cytokine that can stimulate colony formation in bone marrow megakaryocytes. Mesothelin: an anchored cell-surface protein, may function as a cell adhesion protein. Overexpressed in epithelial mesotheliomas, ovarian cancers, some squamous cell carcinomas.	high: lung	sequenced
<b>MSLNL</b>	chr16:759429-772927	mesothelin-like: little known, but may be involved in cell adhesion,	n/a	unlikely



		vacuolar fusion protein MON1		
<b>MIR662</b>	chr16:760184-760278	microRNA 662	n/a	n/a
<b>RPUSD1</b>	chr16:774975-778384	RNA pseudouridylation synthase domain containing 1: involved in pseudouridine synthesis	high: Burkitts lymphoma, B lymphoblasts	unlikely
<b>CHTF18</b>	chr16:778623-788075	chromosome transmission fidelity factor 18 homolog (S. cerevisiae): are components of an alternative replication factor C (RFC) complex that loads PCNA onto DNA during S phase of the cell cycle	high: lymphoblasts	unlikely
<b>GNG13</b>	chr16:788042-790734	guanine nucleotide binding protein: function as signal transducers for the 7-transmembrane-helix G protein-coupled receptors. GNG13 is a gamma subunit that is expressed in taste, retinal, and neuronal tissues and plays a key role in taste transduction	high: thalamus, cerebellum	unlikely
<b>PRR25</b>	chr16:795444-803862	proline rich 25: "A small proline-rich protein, spr1: specific marker for squamous lung carcinoma"	n/a	unlikely
<b>LMF1</b>	chr16:843635-960985	involved in the maturation and transport of lipoprotein lipase through the secretory pathway. Mutations in this gene are associated with combined lipase deficiency	high: adipose tissue, skeletal muscle, heart, and liver.	unlikely
<b>SOX8</b>	chr16:971809-976980	involved in the regulation of embryonic development and in determination of the cell fate. The protein may act as a transcriptional activator after forming a complex with other proteins. may be involved in brain development and function. Haploinsufficiency for this protein may contribute to the mental retardation found in haemoglobin H-related mental retardation (ART-16 syndrome).	high: brain regions, most tissues	unlikely
<b>SSTR5</b>		somatostatin R5: this receptor is mediated by G proteins which inhibit adenylyl cyclase, and different regions of this receptor molecule are required for the activation of different signaling pathways. A mutation in this gene results in somatostatin analog resistance	high: brain, endocrine and nervous system	unlikely
<b>C1QTNF8</b>	chr16:1078227-1086245	C1q and tumor necrosis factor related protein 8	n/a	n/a
<b>CACNA1H</b>	chr16:1143242-1211773	calcium channel, voltage-dependent, T type, alpha 1H subunit: encodes a T-type member of the alpha-1 subunit family. a protein in the voltage-dependent calcium channel complex. Calcium channels mediate the influx of calcium ions into the cell upon membrane polarization	high: adrenal cortex, brain	very unlikely
<b>TPSG1</b>	chr16:1211653-1215255	Trypsins comprise a family of trypsin-like serine proteases, the peptidase family S1. Trypsins are enzymatically active only as heparin-stabilized tetramers, and they are resistant to all known endogenous proteinase inhibitors. Trypsins have been implicated as mediators in the pathogenesis of asthma and other allergic and inflammatory disorders.	high: most human tissues, not in skeletal muscle/ spleen	possible
<b>TPSB2</b>	chr16:1218337-1220186			
<b>TPSAB1</b>	chr16:1230679-1232556			
<b>TPSD1</b>	chr16:1246274-1248495			
<b>UBE2I</b>	chr16:1299181-1315391	ubiquitin-conjugating enzyme E2I (UBC9 homolog, yeast): modification of proteins with ubiquitin is an important cellular mechanism for targeting abnormal or short-lived proteins for degradation.	high: most tissues	unlikely

<b>BAIAP3</b>	chr16:1323607-1339443	BAI1-associated protein 3 (BAIAP3): This p53-target gene encodes a brain-specific angiogenesis inhibitor. Its expression pattern and similarity to other proteins suggest that it may be involved in synaptic functions.	high: pituitary, hypothalamus	very unlikely
<b>GNPTG</b>	chr16:1341901-1353353	N-acetylglucosamine-1-phosphate transferase: catalyzes the first step in synthesis of a mannose 6-phosphate lysosomal recognition marker. Mutations in the gene encoding the gamma subunit have been associated with mucopolidosis IIIC	high: thyroid, amygdala, pineal, lung	unlikely
<b>UNKL</b>	chr16:1353207-1404706	unkempt homolog (Drosophila)-like (UNKL): RING finger protein unkempt-like isoform 1	in most tissues	unlikely
<b>CCDC154</b>	chr16:1424390-1434491	coiled-coil domain containing 154: involved with Tropomyosin	n/a	unlikely
<b>CLCN7</b>	chr16:1434936-1465086	chloride channel 7: Defects in this gene are the cause of osteopetrosis autosomal recessive type 4. Osteopetrosis is a rare genetic disease characterized by abnormally dense bone, due to defective resorption of immature bone. OPTA2 is the most common form of osteopetrosis, occurring in adolescence or adulthood.	high: adrenal cortex/gland, retina, PFC	unlikely
<b>PTX4</b>	chr16:1475941-1478469	pentraxin 4, long: Some PTXs are part of the humoral arm of innate immunity and behave as functional ancestors of antibodies by mediating agglutination, complement activation, and opsonization... <i>affect macrophage activation</i>	high: in small intestine, testis, and bone marrow... low: brain	possible
<b>TELO2</b>	chr16:1483353-1500461	telomere maintenance 2: encodes a protein that functions as an S-phase checkpoint protein in the cell cycle. The protein may also play a role in DNA repair	n/a	sequenced
<b>IFT140</b>	chr16:1500429-1602110	intraflagellar transport 140 homolog (Chlamydomonas); WD40 repeat, Quinoprotein amine dehydrogenase, beta chain-like	high: thyroid	unlikely
<b>TMEM204</b>	chr16:1524232-1545244	transmembrane protein 204: C16ORF30 plays a role in cell adhesion and cellular permeability at adherens junctions	highest: lung	possible
<b>CRAMP1L</b>	chr16:1604642-1667910	Crm, cramped-like (Drosophila): DNA-binding.	n/a	unlikely
<b>HN1L</b>	chr16:1668279-1692074	hematological and neurological expressed 1-like: which are proposed to be involved in embryo development	liver, kidney, prostate, testis and uterus at varying levels	unlikely
<b>MAPK8IP3</b>	chr16:1696222-1760319	shares similarity with the product of Drosophila synd gene, required for interact of kinesin I with axonal cargo. Studies of the similar gene in mouse suggested that this protein may regulate protein kinases of the JNK signaling pathway, thus function as a scaffold protein in neuronal cells. C. elegans counterpart of this gene regulates synaptic vesicle transport	high: brain, endothelial cells	unlikely
<b>microRNA 3177</b>	chr16:1724987-1725068	n/a	n/a	
<b>NME3</b>	chr16:1760322-1761711	non-metastatic cells 3, protein expressed in (NME3): role for the NME3	high: prostate,	unlikely

		gene in normal hematopoiesis and raised the possibility that its overexpression contributes to differentiation arrest, a feature of blastic transformation in chronic myelogenous leukemia	pineal gland, heart... lung: 4xM	
<b>MRPS34</b>	chr16:1761897-1763141	mitochondrial ribosomal protein S34 (MRPS34), nuclear gene encoding mitochondrial protein	high: burkitts lymphoma, B lymphoblasts	unlikely
<b>EME2</b>	chr16:1763230-1766240	ME2 forms a heterodimer with MUS81 that functions as an XPF-type flap/fork endonuclease in DNA repair (OMIM)	n/a	unlikely
<b>SPSB3</b>	chr16:1766714-1772582	spiA/ryanodine receptor domain and SOCS box containing 3:	high: CD4/CD8 T cells	possible but unlikely
<b>NUBP2</b>	chr16:1772934-1779193	NUBP2 is a member of the NUBP/MRP gene subfamily of ATP-binding proteins:	highest: heart, subthalamic nucleus, endothelial cells... also high in lung	unlikely
<b>IGFALS</b>	chr16:1780415-1783735	insulin-like growth factor binding protein, acid labile subunit (IGFALS): serum protein that binds insulin-like growth factors, increasing their half-life and vascular localization. Production of this protein is stimulated by growth hormone. Defects in this gene are a cause of acid-labile subunit deficiency, which manifests itself in a delayed and slow puberty.	high: liver	very unlikely
<b>HAGH</b>	chr16:1799105-1817196	hydroxyacylglutathione hydrolase: a thioesterase and is responsible for the hydrolysis of S-lactoyl-glutathione to reduced glutathione and D-lactate.	high: liver, early erythroid	very unlikely
<b>FAHD1</b>	chr16:1817226-1830204	fumarylacetoacetate hydrolase domain containing 1 (FAHD1), nuclear gene encoding mitochondrial protein:	n/a	unlikely
<b>C16orf73</b>	chr16:1823985-1862180	hypothetical protein LOC254528 isoform 1	n/a	n/a
<b>NCRNA00254</b>	chr16:1868287-1874233	non-protein coding RNA 254	n/a	n/a
<b>HS3ST6</b>	chr16:1901466-1908232	heparan sulfate (glucosamine) 3-O-sulfotransferase 6: binds to HSV-1 (herpes simplex virus type 1), the resulting pr serves as a entry R for HSV-1	high: mucosal membranes	unlikely
<b>SEPX1</b>	chr16:1928235-1933295	selenoprotein X, 1: This protein belongs to the methionine sulfoxide reductase B (MsrB) family, and it is expressed in a variety of adult and fetal tissues, contains a selenocysteine (Sec) residue at its active site	high: liver, leukocytes... low: lung, brain	unlikely
<b>RPL3L</b>	chr16:1934581-1944680	ribosomal protein L3-like: The protein belongs to the L3P family of ribosomal proteins. A novel ribosomal protein L3-like gene (RPL3L) maps to the autosomal dominant polycystic kidney disease gene region	high: skeletal muscle, heart	unlikely
<b>NDUFB10</b>	chr16:1949518-1951977	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 10: part of the first enzyme complex in the electron transport chain of mitochondria.	high: heart	unlikely
<b>RPS2</b>	chr16:1952063-1954828	ribosomal protein S2: This gene encodes a ribosomal protein that is a	high: CD cells	unlikely

		component of the 40S subunit. The protein belongs to the S5P family of ribosomal proteins. It is located in the cytoplasm.... Role in controlling let-7a expression in human prostate cancer		
<b>SNORA10</b>	chr16:1952336-1952468	small nucleolar RNA, H/ACA box 10	n/a	n/a
<b>SNORA64</b>	chr16:1952975-1953108	small nucleolar RNA, H/ACA box 64		
<b>SNHG9</b>	chr16:1954998-1955506	small nucleolar RNA host gene 9 (non-protein coding)	n/a	unlikely
<b>SNORA78</b>	chr16:1955186-1955312	small nucleolar RNA, H/ACA box 78	n/a	n/a
<b>RNF151</b>	chr16:1956876-1958977	ring finger protein 151: RNF151, a testis-specific RING finger protein, interacts with dysbindin... may play a role in acrosome formation	high: testis	very unlikely
<b>TBL3</b>	chr16:1962065-1968752	transducin (beta)-like 3: sequence similarity with members of the WD40 repeat-containing protein family. The WD40 group = large family of proteins, appear to have a regulatory function.. believed that WD40 repeats mediate pr-pr interactions could be involved in signal transduction, RNA processing, gene regulation, vesicular trafficking, cytoskeletal assembly and maybe control of cytotypic differentiation. This gene has multiple polyadenylation sites.	high: lung, testis, liver, heart, prostate, thyroid, lymphoblasts	possible but unlikely
<b>NOXO1</b>	chr16:1969042-1971185	NADPH oxidase organizer 1: catalyze the transfer of electrons from NADPH to molecular oxygen to generate reactive oxygen species (ROS).	High: colon, liver, kidneys	possible (NOXO1 and lung, and fibrosis)
<b>GFER</b>	chr16:1974151-1977751	growth factor, augments of liver regeneration: hepatotropic factor designated augment of liver regeneration (ALR) is thought to be one factor responsible for the extraordinary regenerative capacity of mammalian liver... Mutation = autos recessive myopathy	high: kidneys, brain	unlikely
<b>SYNGR3</b>	chr16:1979947-1984277	synaptogyrin 3: encodes an integral membrane protein. The exact function of this protein is unclear, but studies of a similar murine protein suggest that it is a synaptic vesicle protein that also interacts with the dopamine transporter	brain and placenta only	unlikely
<b>ZNF598</b>	chr16:1987769-1999764	zinc finger protein 598:	n/a	unlikely
<b>NPW</b>	chr16:2009522-2010757	neuropeptide W: 23- and 30-amino acid neuropeptides that bind/activate two G-protein coupled Rs in the CNS. The neuropeptides enhance cortisol secretion from adrenal cells via adenylylate cyclase/PKA signaling cascade.	high: brain, lymphoblastic leukemia, fetal kidney, colorectal adenocarcinoma, trachea	unlikely
<b>SLC9A3R2</b>	chr16:2016889-2029028	solute carrier family 9 (sodium/hydrogen exchanger), member 3 regulator 2: is a sodium/hydrogen exchanger in brush border membrane of the proximal tubule, SI and colon- plays a major role in transepithelial sodium absorption	high: adrenal cortex, liver, skeletal muscle,	unlikely
<b>NTHL1</b>	chr16:2029817-2037868	nth endonuclease III-like 1 (E. coli): the encoded protein has DNA glycosylase activity on DNA substrates containing oxidized pyrimidine	high: liver, heart.. lung: over	possible

		residues and has apurinic/aprimidinic lyase activity... <i>linked to COPD, lung cancer</i>	3xM	
<b>TSC2</b>	chr16:2037991-2078714	tuberous sclerosis 2: Mutations in this gene lead to tuberous sclerosis complex. Its gene product is believed to be a tumor suppressor and is able to stimulate specific GTPases. The protein associates with hamartin in a cytosolic complex, possibly acting as a chaperone for hamartin.... <i>mutations also causal in Pulmonary Lymphangioleiomyomatosis</i>	high: thyroid, heart, pineal	unlikely
<b>PKD1</b>	chr16:2078712-2125900	polycystic kidney disease 1 (autosomal dominant): mutations in this gene cause autosomal dominant polycystic kidney disease type 1, characterized by the growth of fluid-filled cysts that replace normal renal tissue and result in end-stage renal failure	high: renal/epithelial cells	unlikely
<b>MIR1225</b>	chr16:2080197-2080286	microRNA 1225: belongs to subfamily of miRNAs called mirtrons, that arise from short hairpin introns. Mirtrons differ from in that they use splicing rather than cleavage by Drosha	n/a	unlikely
<b>MIR3180-5</b>	chr16:2125979-2126131	microRNA 3180-5	n/a	unlikely
<b>RAB26</b>	chr16:2138652-2144142	member RAS oncogene family: regulators of vesicular fusion/trafficking. This G-prtn family regulates intercellular vesicle trafficking, including exocytosis, endocytosis, recycling	high: adult/fetal brain	unlikely
<b>SNORD60</b>	chr16:2145025-2145107	small nucleolar RNA, C/D box 60	n/a	n/a
<b>TRAF7</b>	chr16:2145800-2168131	TNF receptor-associated factor 7: signal transducers for members of the TNF receptor superfamily	high: colorectal adenocarcinoma, fetal brain, B lymphoblasts	unlikely
<b>CASKIN1</b>	chr16:2167185-2186466	CASK interacting protein 1: coat the cytoplasmic tails of neuroligins and other cell surface proteins.	high: brain, lung, liver	unlikely
<b>MLST8</b>	chr16:2195179-2199419	MTOR associated protein, LST8 homolog (S. cerevisiae): regulates cell growth in response to nutrients and growth factors.	high: prostate, liver	unlikely
<b>C16orf79</b>	chr16:2199255-2201070	n/a	n/a	n/a
<b>PGP</b>	chr16:2201604-2204823	phosphoglycolate phosphatase: may have an important regulatory influence on oxygen transport in man	high: skeletal, cardiac tissues	unlikely
<b>E4F1</b>	chr16:2213568-2225744	E4F transcription factor 1: The zinc finger protein is one cell transcription factors whose DNA-binding activities are regulated through adenovirus E1A. plays an important role in the cellular life-or-death decision controlled by p53.	high: CD4, CD8 T cells	unlikely
<b>DNASE1L2</b>	chr16:2226469-2228713	deoxyribonuclease I-like 2: strong endonuclease activity in the presence of both Ca(2+) and Mg(2+). Inhibited by increasing salt concentration	high: brain, pituitary	unlikely
<b>DCI</b>	chr16:2229874-2241603	dodecenoyl-CoA isomerase (DCI), nuclear gene encoding mitochondrial protein: a key mitochondrial enzyme involved in beta-oxidation of unsaturated FA's	high: liver, heart	very unlikely
<b>RNPS1</b>	chr16:2243101-2257859	RNA binding protein S1, serine-rich domain: protein binds to the mRNA and remains bound after nuclear export, acting as a	high: prostate	unlikely

		nucleocytoplasmic shuttling protein...strongly activate splicing of constitutively and alternatively spliced pre-mRNAs		
<b>MIR3677</b>	chr16:2260715-2260774		n/a	n/a
<b>MIR940</b>	chr16:2261749-2261842		n/a	n/a
<b>ABCA3</b>	chr16:2265880-2330748	ATP-binding cassette, sub-family A (ABC1), member 3: ABC proteins transport various molecules across extra- and intracellular membranes... have been linked with IPF "Usual interstitial pneumonia in an adolescent with ABCA3 mutations" and other papers.	high: lung	sequenced (Fady)
<b>ABCA17P</b>	chr16:2330924-2416701	ATP-binding cassette, sub-family A (ABC1), member 17, pseudogene: shares a common 5' end with ABCA3	testis-specific	unlikely
<b>CCNF</b>	chr16:2419396-2448860	cyclin F: important regulators of cell cycle transitions through their ability to bind and activate cyclin-dependent protein kinases... the F-box proteins function in phosphorylation-dependent ubiquitination.	high: early erythroid	unlikely
<b>C16orf59</b>	chr16:2450116-2454965		n/a	n/a
<b>NTN3</b>	chr16:2461501-2464147	netrin 3: netrins are a family of chemotropic factors that guide axons to their targets	high: ciliary/dorsal root ganglions	very unlikely
<b>TBC1D24</b>	chr16:2465148-2495735	TBC1 domain family, member 24: suggesting it is involved in cell signaling: is mutated in familial infantile myoclonic epilepsy	high: brain, followed by testis, skeletal muscle, heart, kidney, lung, and liver	unlikely
<b>ATP6V0C</b>	chr16:2503728-2510225	ATPase, H <sup>+</sup> transporting, lysosomal 16kDa, V0 subunit c: a multisubunit enzyme that mediates acidification of eukaryotic intracellular organelles, which is necessary for protein sorting, zymogen activation, receptor-mediated endocytosis, synaptic vesicle proton gradient generation	high: whole brain, lung	unlikely?
<b>AMDHD2</b>	chr16:2510364-2519733	amidohydrolase domain containing 2: product is putative N-acetylglucosamine-6-phosphate	highest in dendritic cells	unlikely
<b>CEMP1</b>	chr16:2520037-2521410	cementum protein 1: could be involved in regulating the cementogenesis process during root development.	high: periodontal ligament/ cementum-forming cells	unlikely
<b>MIR3178</b>	chr16:2521924-2522007		n/a	
<b>PDPK1</b>	chr16:2527971-2593190	3-phosphoinositide dependent protein kinase-1: activates PKB, Isoforms of protein kinase B are overexpressed in some ovarian, pancreatic, and breast cancer cells, and PKB has been shown to protect cells from apoptosis.	high: prostate, CD34/CD33 cells	unlikely
<b>LOC652276</b>	chr16:2593386-2620496	hypothetical LOC652276 (LOC652276), non-coding RNA		

<b>FLJ42627</b>	chr16:2628984-2636131	hypothetical LOC645644 (FLJ42627), non-coding RNA		
<b>KCTD5</b>	chr16:2672496-2699032	potassium channel tetramerisation domain containing 5: interacts with the large regulatory proteins Rep78, Rep68 of adeno-associated virus-2 (AAV-2).	high: bronchial epithelial cells, lung	unlikely
<b>PRSS27</b>	chr16:2702424-2710553	protease, serine 27: Pancreasin is a pancreatic tryptic serine peptidase that cleaves peptides after an arginine residue.	high: pancreas, lung	possible but unlikely
<b>LOC100128788</b>	chr16:2727078-2742602	hypothetical LOC100128788		
<b>SRRM2</b>	chr16:2742331-2761414	serine/arginine repetitive matrix 2: a potential blood biomarker revealing high alternative splicing in Parkinson's disease.	expressed in all tissues	unlikely
<b>TCEB2</b>	chr16:2761416-2767298	transcription elongation factor B (SIII), polypeptide 2 (elongin B): encodes the protein elongin B, which is a subunit of the transcription factor B (SIII) complex. the tumor suppression activity of the von Hippel-Lindau (608537) tumor suppressor gene product is a function of its ability to bind to elongins B and C, thus inhibit transcription elongation... <i>The von Hippel-Lindau Chuvash mutation promotes pulmonary hypertension and fibrosis in mice (paper).</i>	high: liver, heart, bronchial epithelial cells, testis, pineal	possible
<b>PRSS33</b>	chr16:2773955-2776709	protease, serine, 33	n/a	n/a
<b>PRSS41</b>	chr16:2788487-2795134	protease, serine, 41	n/a	n/a
<b>PRSS21</b>	chr16:2807165-2811719	testisin: encodes a cell-surface anchored serine protease, a member of the trypsin family of serine proteases. It may be involved in progression of testicular tumors of germ cell origin.	high: testicular cells	unlikely
<b>ZG16B</b>	chr16:2820174-2822286	zymogen granule protein 16 homolog B (rat):	high: salivary gland, trachea	unlikely
<b>PRSS30P</b>	chr16:2829575-2832753	protease, serine, 30 homolog (mouse), pseudogene, non-coding RNA	n/a	n/a
<b>PRSS22</b>	chr16:2842729-2848172	This gene encodes a member of the trypsin family of serine proteases. The enzyme is expressed in the airways in a developmentally regulated manner	high: esophagus, trachea.... lung: low	unlikely
<b>FLYWCH2</b>	chr16:2873197-2889384	FLYWCH family member 2	high: colorectal adenocarcinoma	unlikely
<b>FLYWCH1</b>	chr16:2901981-2930159	FLYWCH-type zinc finger-containing protein 1	high: CD4 T cells, thalamus	unlikely
<b>KREMEN2</b>	chr16:2954218-2958382	kringle containing transmembrane protein 2: The encoded protein forms a ternary membrane complex with DKK1 and WNT receptor lipoprotein receptor-related protein 6 (LRP6), inducing rapid endocytosis and removal of LRP6 from the plasma membrane.. <i>Inhibition of Wnt/beta-catenin/CREB binding protein (CBP) signaling reverses pulmonary fibrosis... Dickkopf proteins influence lung epithelial cell proliferation in idiopathic pulmonary fibrosis (paper).</i>	high in liver	possible
<b>PAQR4</b>	chr16:2959343-2963486	progesterone and adipoQ receptor family member IV: hormones bind to	high: retina	unlikely

		these Rs, involved in a variety of metabolic processes		
<b>PKMYT1</b>	chr16:2962793-2970541	protein kinase, membrane associated tyrosine/threonine 1: This kinase preferentially phosphorylates and inactivates cell division cycle 2 protein (CDC2), and thus negatively regulates cell cycle G2/M transition. This kinase is associated with the membrane throughout the cell cycle.	high: early erythroid	unlikely
<b>LOC283875</b>	chr16:2979056-2984511	hypothetical LOC283875	n/a	n/a
<b>CLDN9</b>	chr16:3002458-3004507	Claudins are integral membrane proteins and components of tight junction strands. This protein is one of the entry cofactors for hepatitis C virus. Mouse studies showed that this gene is needed for the preservation of sensory cells in the hearing organ and the gene deficiency is associated with deafness.	high: heart, skeletal muscle	unlikely
<b>CLDN6</b>	chr16:3004714-3008189			
<b>TNFRSF12A</b>	chr16:3010314-3012384	tumor necrosis factor receptor superfamily, member 12A: stimulation of cell growth and angiogenesis, induction of inflammatory cytokines, and under some experimental conditions, stimulation of apoptosis....	high: heart, kidney, bronchial epithelial cells... lung: intermediate expression	possible
<b>HCFC1R1</b>	chr16:3012627-3014288	host cell factor C1 regulator 1 (XPO1 dependent): the HCFC1 protein is involved in control of the cell cycle and transcriptional regulation during herpes simplex virus infection	high: whole brain, heart, lung	unlikely
<b>THOC6</b>	chr16:3014033-3017757	THO complex 6 homolog (Drosophila): THO complex: nuclear complex that is required for transcription elongation through genes containing tandemly repeated DNA seq's.	high: CD cells	unlikely
<b>CCDC64B</b>	chr16:3017869-3025543	coiled-coil domain containing 64B: molecular function- Rab GTPase binding	n/a	n/a
<b>MMP25</b>	chr16:3036683-3050725	matrix metalloproteinase 25: involved in breakdown of ECM in normal physiological processes (embryonic development, reproduction, tissue remodeling), also in disease processes (arthritis and metastasis).	high: leukocytes, lung, and spleen	possible but unlikely
<b>IL32</b>	chr16:3055314-3059669	IL32 (a cytokine) stimulates monocyte cells to produce various proinflammatory cytokines and chemokines through the NF-kappa-B and p38 MAPK inflammatory signal pathway... may be involved in inflammatory/autoimmune diseases... associated with COPD.	high: CD4/8 T cells, B lymphoblasts	possible
<b>ZSCAN10</b>	chr16:3078896-3082862	zinc finger and SCAN domain containing 10:	n/a	n/a
<b>MGC3771</b>	chr16:3100462-3105600	hypothetical LOC81854 (MGC3771), transcript variant 1, non-coding RNA	high: cardiac myocytes, skeletal muscle	unlikely
<b>ZNF205</b>	chr16:3102564-3110519	Kruppel-type zinc finger protein that contains a KRAB A box N-terminal to 8 zinc finger motifs.	high: cardiac myocytes, some brain regions	unlikely
<b>ZNF213</b>	chr16:3125058-3132806	C2H2 zinc finger proteins such as ZNF213: have bipartite structures in which one domain binds DNA/RNA and the other modulates target gene expression	high: testis, ovary	unlikely



<b>OR1F1</b>	chr16:3194248-3195186	Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell. The olfactory receptor proteins are members of a large family of G-protein-coupled receptors (GPCR) arising from single coding-exon genes.	high: heart, liver, ganglion cells, some brain regions	unlikely
<b>OR1F2P</b>	chr16:3205563-3206547			
<b>ZNF200</b>	chr16:3212326-3225187	zinc finger protein 200: novel binding partner of histone H3 methyltransferase G9a.	high: testis	unlikely
<b>MEFV</b>	chr16:3232029-3246628	encodes a protein: pyrin or marenostrin: is an important modulator of innate immunity. Mutations in this gene are associated with Mediterranean fever, a hereditary periodic fever syndrome.	high: ganglion cells, peripheral blood,	unlikely
<b>FLJ39639</b>	chr16:3253769-3257567	hypothetical protein FLJ39639	n/a	n/a
<b>ZNF263</b>	chr16:3273488-3281460	zinc finger protein 263	high: retina	unlikely
<b>TIGD7</b>	chr16:3288809-3295440	tigger transposable element derived 7: tigger subfamily of pogo superfamily of DNA-mediated transposons in humans. These proteins: related to DNA transposons found in fungi/nematodes, and more distantly to the Tc1/mariner transposases... also very similar to the major mammalian centromere protein B... exact function of this gene is not known.	high: ciliary ganglion, atrioventricular node	unlikely
<b>ZNF75A</b>	chr16:3295434-3308577	zinc finger protein 75a:	ubiquitous exp	unlikely
<b>OR2C1</b>	chr16:3345890-3346925	olfactory receptor, family 2, subfamily C, member 1: function above in "OR1F1"	high: heart, liver, ganglion cells, some brain regions	unlikely
<b>MTRNR2L4</b>	chr16:3361054-3362284	MT-RNR2-like 4: function not well known, in extracellular region	n/a	unlikely
<b>ZNF434</b>	chr16:3372086-3391026	zinc finger protein 434	n/a	unlikely
<b>ZNF174</b>	chr16:3391191-3395044	zinc finger protein 174- isoform b		
<b>ZNF597</b>	chr16:3426111-3433491	zinc finger protein 597		
<b>NAT15</b>	chr16:3447993-3476964	N-acetyltransferase 15 (GCN5-related, putative) (NAT15), transcript variant 3: SNPs have been associated with lung cancer.	high: lung, liver, heart	possible but unlikely
<b>C16orf90</b>	chr16:3483485-3485422	chromosome 16 open reading frame 90: predicted	n/a	unlikely
<b>CLUAP1</b>	chr16:3490964-3526586	clusterin associated protein 1: expression is frequently upregulated in colon cancer, provisional status	high: pineal	unlikely
<b>NLRC3</b>	chr16:3529037-3567393	NLR family, CARD domain containing 3:	n/a	unlikely
<b>SLX4</b>	chr16:3571184-3601586	SLX4 structure-specific endonuclease subunit homolog (S. cerevisiae): encodes a structure-specific endonuclease subunit. The encoded protein contains a central BTB domain and it forms a multiprotein complex. is required for repair of specific types of DNA lesions and is critical for cellular responses to replication fork failure. The encoded protein acts as a docking platform for the assembly of multiple, mutated in FANCONI ANEMIA	high: brain, lung, thyroid	unlikely
<b>DNASE1</b>	chr16:3642941-3648097	deoxyribonuclease 1: encodes a member of DNase family. This protein	high: small	unlikely

		is stored in zymogen granules of the nuclear envelope/functions by cleaving DNA. Mutations in gene: associated with systemic lupus erythematosus (SLE), an autoimmune disease. A recombinant form of this protein is used to treat the one of the symptoms of cystic fibrosis by hydrolyzing the extracellular DNA in sputum and reducing its viscosity.	intestine, pancreas	
TRAP1	chr16:3648039-3707599	TNF receptor-associated protein 1: HSP90 proteins are highly conserved molecular chaperones that have key roles in signal transduction, protein folding, protein degradation, and morphologic evolution.	high: localizes to mitochondria	unlikely
CREBBP	chr16:3715057-3870122	CREB binding protein, transcript variant 2: involved in transcriptional coactivation of many different TFs. First isolated as a nuclear protein that binds to cAMP-response element binding protein (CREB), this gene is now known to play critical roles in embryonic development, growth control, and homeostasis by coupling chromatin remodeling to transcription factor recognition.	ubiquitous exp	unlikely
ADCY9	chr16:3952651-4106187	adenylate cyclase 9: catalyses the formation of cyclic AMP from ATP: type 9 is a widely distributed adenylyl cyclase, and it is stimulated by beta-adrenergic receptor activation but is insensitive to forskolin, calcium, and somatostatin	high: colon.... moderate: lung, kidneys, spleen, testis, ovary	unlikely
SRL	chr16:4179376-4232082	sarcalumenin: this gene encodes a 160-kD glycoprotein protein involved in calcium signaling.	high: muscle cells	unlikely
TFAP4	chr16:4247188-4263002	transcription factor AP-4 (activating enhancer binding protein 4): TFAP4 activates both viral and cellular genes by binding to the symmetrical DNA sequence CAGCTG; target of MYC, blocks p21 expression (overexpression is a prognostic indicator for gastric carcinoma)	high: colorectal carcinoma cells	unlikely
GLIS2	chr16:4322226-4329599	GLIS family zinc finger 2: function as activators and/or repressors of gene transcription; plays an essential role in the maintenance of renal tissue architecture through prevention of apoptosis and fibrosis.	high: kidney.... low: heart, lung, placenta	sequenced (Fady)
CORO7-PAM16	chr16:4330253-4406963	the conjoined read-through of 2 genes (listed below)		
PAM16	chr16:4330253-4341374	presequene translocase-associated motor 16 homolog (S. cerevisiae) (PAM16), nuclear gene encoding mitochondrial protein	high: brain, thymus, kidney	unlikely
CORO7	chr16:4344544-4406963	coronin 7: plays a role in Golgi complex morphology and function		
VASN	chr16:4361850-4373530	vasorin: directly binds to transforming growth factor (TGF)-beta and attenuates TGF-beta signaling; potential therapeutic target for vascular fibroproliferative disorders by modulating arterial response	high: vascular smooth muscle cells	sequenced (Fady)
DNAJA3	chr16:4415859-4446776	DnaJ (Hsp40) homolog, subfamily A, member 3: belongs to the evolutionarily conserved DNAJ/HSP40 family of proteins;	high: heart, liver, skeletal muscle	unlikely
HMOX2	chr16:4466342-4500349	heme oxygenase (decycling) 2: part of the heme degradative pathway:	high: testis,	unlikely

		cleaves heme to form biliverdin, which is subsequently converted to bilirubin by biliverdin reductase, and carbon monoxide, a putative neurotransmitter.	whole brain, brochial epithelial cells, lung, liver, heart.	
<b>C16orf5</b>	chr16:4500678-4528817	chromosome 16 open reading frame 5: ascertained in a patient with epilepsy and mental retardation	high: brain.... Little to no expression in lung	very unlikely
<b>FAM100A</b>	chr16:4598885-4604928	family with sequence similarity 100, member A: hypothetical protein LOC124402	high: thyroid, placenta	unlikely
<b>MGRN1</b>	chr16:4614827-4680976	mahogunin, ring finger 1: is a C3HC4 RING-containing protein with E3 ubiquitin ligase activity in vitro. Mahogunin is the protein mutant in 'mahoganoid,' a coat-color mutation of mice in which pigmentation is very similar to that of 'mahogany.'	high: brain, lung, liver	unlikely
<b>NUDT16L1</b>	chr16:4683695-4685861	nudix (nucleoside diphosphate linked moiety X)-type motif 16-like 1	high: heart, prostate, thyroid, liver	unlikely
<b>C16orf71</b>	chr16:4724290-4739398	chromosome 16 open reading frame 71: hypothetical protein LOC146562	high: testis	very unlikely
<b>ZNF500</b>	chr16:4740816-4757167	zinc finger protein 500	n/a	unlikely
<b>SEPT.12</b>	chr16:4767673-4778400	septin 12, trans. var. 2: conserved GTP-binding proteins that function as dynamic, regulatable scaffolds for recruitment of other proteins/involved in membrane dynamics, vesicle trafficking, apoptosis, and cytoskeleton remodeling, as well as infection, neurodegeneration, and neoplasia.	high: testis	unlikely
<b>GLYR1</b>	chr16:4793205-4837304	glyoxylate reductase 1 homolog (Arabidopsis): localizes to the nucleus in a variety of cell lines	high: CD cells	unlikely
<b>UBN1</b>	chr16:4837913-4872364	ubiquitin 1, trans var 2: in nucleus, interacts with some cellular/viral TFs	ubiquitous exp	unlikely
<b>PPL</b>	chr16:4872509-4927137	periplakin: a component of desmosomes and of the epidermal cornified envelope in keratinocytes. its C-terminus interacts with intermediate filaments. "Identification of Periplakin as a New Target for Autoreactivity in Idiopathic Pulmonary Fibrosis" <i>Am J Respir Crit Care Med.</i>	high: epidermis, stomach, cornea	sequenced (Fady)
<b>SEC14L5</b>	chr16:4948319-5009157	SEC14-like 5 (S. cerevisiae)	high: OFB, pineal	unlikely
<b>NAGPA</b>	chr16:5014846-5023943	N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase: This gene encodes the enzyme that catalyzes the second step in the formation of the mannose 6-phosphate recognition marker on lysosomal hydrolases.	high: brain, myeloid cells	unlikely
<b>C16orf89</b>	chr16:5034124-5056112	chromosome 16 open reading frame 89: hypothetical protein LOC146556	high: thyroid, lung	possible but unlikely
<b>ALG1</b>	chr16:5061811-5077381	asparagine-linked glycosylation 1, beta-1,4-mannosyltransferase	n/a	unlikely

		homolog ( <i>S. cerevisiae</i> ): enzyme encoded by this gene catalyzes the first mannosylation step in the biosynthesis of lipid-linked oligosaccharides. This gene is mutated in congenital disorder of glycosylation type Ik.		
<b>FAM86A</b>	chr16:5074302-5087790	family with sequence similarity 86, member A: hypothetical protein LOC196483 isoform 1	high: pituitary, liver	unlikely
<b>RBFOX1</b>	chr16:6009133-7703341	RNA binding protein, fox-1 homolog ( <i>C. elegans</i> ) 1: Ataxin-2 binding protein 1 has an RNP motif that is highly conserved among RNA-binding proteins. This protein binds to the C-terminus of ataxin-2 and may contribute to the restricted pathology of spinocerebellar ataxia type 2	high: whole brain	unlikely
<b>TMEM114</b>	chr16:8559503-8562227	transmembrane protein 114	high: eye, brain, testis	unlikely
<b>METTL22</b>	chr16:8623028-8647580	methyltransferase like 22: hypothetical protein LOC79091	n/a	unlikely
<b>ABAT</b>	chr16:8722074-8785933	4-aminobutyrate aminotransferase: responsible for catabolism of gamma-aminobutyric acid (GABA, a NT), into succinic semialdehyde. BAT deficiency phenotype includes psychomotor retardation, hypotonia, hyperreflexia, lethargy, refractory seizures, and EEG abnormalities.	high: brain, pituitary, liver	unlikely
<b>TMEM186</b>	chr16:8796538-8799006	transmembrane protein 186	high: B lymphoblasts	unlikely
<b>PMM2</b>	chr16:8799171-8850695	phosphomannomutase 2: catalyzes the isomerization of mannose 6-P to mannose 1-P. Mutations in PMM2: cause defects in glycoprotein biosynthesis, which manifests as carbohydrate-deficient glycoprotein syndrome type I.	high: CD cells, B lymphoblasts, small intestine, smooth muscle	very unlikely
<b>CARHSP1</b>	chr16:8854303-8870364	calcium regulated heat stable protein 1, 24kDa: is required for effective tumor necrosis factor alpha mRNA stabilization	high: testis,.... Moderate: liver, lung, placenta	unlikely
<b>USP7</b>	chr16:8893452-8964842	ubiquitin specific peptidase 7 (herpes virus-associated): may function as a tumor suppressor in vivo through the stabilization of p53... PML-HAUSP network controls PTEN deubiquitinylation and subcellular localization, which is perturbed in human cancers	high: CD 71 early erythroid	possible but unlikely
<b>C16orf72</b>	chr16:9093038-9121056	chromosome 16 open reading frame 72: hypothetical protein LOC29035	n/a	unlikely
<b>MIR548X</b>	chr16:9236275-9236304	microRNAs are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNAs.....	n/a	unlikely
<b>GRIN2A</b>	chr16:9754766-10183764	glutamate receptor, ionotropic, NMDA: class of ionotropic glutamate-gated ion channels; shown to be involved in long-term potentiation (activity-dependent increase in the efficiency of synaptic transmission thought to underlie certain kinds of memory and learning)	high: subthalamic nucleus	unlikely
<b>ATF7IP2</b>	chr16:10430226-10484996	activating transcription factor 7 interacting protein 2: binds to MBD1 which is involved in transcriptional repression and heterochromatin	high: testis, T cells	unlikely

## formation

<b>EMP2</b>	chr16:10529780-10582040	epithelial membrane protein 2: involved in cell proliferation and cell-cell interactions	high: lung, ovary, heart, intestine	sequenced (Fady)
<b>TEKT5</b>	chr16:10628862-10696303	tektin 5 (provisional status): suggests that TEKT5 plays an important role in flagella formation during spermiogenesis/ implicated in sperm motility.	n/a	unlikely
<b>NUBP1</b>	chr16:10745199-10770709	nucleotide binding protein 1 (MinD homolog, E. coli): plays a vital role in cell function	high: lung, testis	unlikely
<b>FAM18A</b>	chr16:10768034-10820122	family with sequence similarity 18, member A (predicted status):	n/a	n/a
<b>CIITA</b>	chr16:10878556-10926341	class II, major histocompatibility complex, transactivator: "Expression of major histocompatibility complex class II antigens on lungs in rat with bleomycin-induced pulmonary fibrosis"	high: liver, skeletal muscle	possible but unlikely
<b>DEXI</b>	chr16:10930249-10943758	Dexi homolog (mouse): upregulated in emphysematous tissue	high: heart, liver, lung	sequenced (Fady)
<b>CLEC16A</b>	chr16:10945846-11183547	C-type lectin domain family 16, member A: associated with Type 1 diabetes	high: kidneys, testis, ovary	unlikely
<b>SOCS1</b>	chr16:11255775-11257540	suppressor of cytokine signaling 1: family members are cytokine-inducible negative regulators of cytokine signaling	high: T cells	sequenced
<b>TNP2</b>	chr16:11269215-11270661	transition protein 2 (during histone to protamine replacement); mammalian spermiogenesis, importin alpha-4 required for active transport of Tnp2 into nuclei of rat testis germ cells	high: testis	very unlikely
<b>PRM3, 2, 1</b>	chr16:11274645-11274953	protamine family: Protamines are the major DNA-binding proteins in the nucleus of sperm, and package the DNA in a volume less than 5% of a somatic cell nucleus	high: testis	unlikely
<b>C16ORF75 aka RMI2</b>	chr16:11346812-11353118	RMI2 is a component of the BLM complex, which plays a role in homologous recombination-dependent DNA repair and is essential for genome stability... involved in Bloom Syndrome	high: lymphoma, lymphoblasts	unlikely
<b>LITAF</b>	chr16:11549079-11588823	lipopolysaccharide-induced TNF factor: Lipopolysaccharide is a potent stimulator of monocytes and macrophages, causing secretion of inflammatory mediators. encodes lipopolysaccharide-induced TNF-alpha factor, a DNA-binding protein and can mediate the TNF-alpha expression by direct binding to the promoter region of the TNF-alpha gene. The transcription of this gene is induced by tumor suppressor p53	highest in lymphoblasts, T cells	possible but unlikely

and has been implicated in the p53-induced apoptotic pathway. Mutations in this gene cause Charcot-Marie-Tooth disease type 1C (CMT1C) and may be involved in the carcinogenesis of extramammary Paget's disease (EMPD). Multiple alternatively spliced transcript variants have been found for this gene.

<b>SNN</b>	chr16:11669790-11680516	stannin: cytokines induced expression of stannin in cultured human umbilical vein endothelial cells...	highest: brain regions	unlikely
<b>TXNDC11</b>	chr16:11680444-11744149	thioredoxin domain containing 11: cell redox homeostasis, sometimes part of the membrane....	high: dendritic cells	unlikely
<b>ZC3H7A</b>	chr16:11751943-11798615	zinc finger CCCH-type containing 7A: functional module in macrophage activation,	low: lung, testis, liver, heart	very unlikely
<b>BCAR4</b>	chr16:11821193-11830190	breast cancer anti-estrogen resistance 4, non-coding RNA	n/a	very unlikely
<b>RSL1D1</b>	chr16:11835556-11852943	ribosomal L1 domain containing 1: RNA binding-structural constituent of ribosome	low: lung, colon, SI	very unlikely
<b>GSPT1</b>	chr16:11869486-11917326	G1 to S phase transition 1, transcript variant 2, mRNA: hypothesized that Gspt1, in a binary complex with eRF1, functions as a polypeptide chain release factor	high: bronchial epithelial cells	unlikely
<b>TNFRSF17</b>	chr16:11966465-11969426	tumor necrosis factor receptor superfamily, member 17: R expressed in mature B lymphocytes, and may be important for B cell development and autoimmune response....may transduce signals for cell survival and proliferation	high: lymphoblasts, dendritic cells	unlikely
<b>RUNDC2A</b>	chr16:11978103-12054642	RUN domain containing 2A:	n/a	unlikely
<b>SNX29</b>	chr16:12053556-12575647	sorting nexin 29: are a large group of proteins that are localized in the cytoplasm and have the potential for membrane association either through lipid-binding PX domain or through protein-protein interactions with membrane-associated protein complexes. Some members of this family have been shown to facilitate protein sorting	ubiquitous	unlikely
<b>CPPED1</b>	chr16:12661157-12805245	calcineurin-like phosphoesterase domain containing 1: hydrolase activity, metal ion binding.	high: CD33 myeloid, CD14 monocytes, cardiac myocytes	unlikely

<b>SHISA9</b>	chr16:12902978-13241774	shisa homolog 9: modulates short-term plasticity at specific excitatory synapses	high: brain (dentate gyrus granule cells)	unlikely
<b>ERCC4</b>	chr16:13921515-13953706	excision repair cross-complementing rodent repair deficiency, complementation group 4: involved in the 5' incision made during nucleotide excision repair..... Defects in this gene are a cause of xeroderma pigmentosum 6	ubiquitous	possible: have been linked to lung cancer...
<b>MKL2</b>	chr16:14072697-14268131	MYOCARDIN-LIKE 2: RNA interference of MKL1 or MKL2 reduced serum and RhoA activation in HeLa cells, and the combination of MKL1 and MKL2 RNA interference completely abolished induction	ubiquitous	unlikely

Gene	Position	Function	Expression	Candidate ?
<b>LOC732275</b>	chr16:84922957-84936786	Homo sapiens similar to hCG1645603 (LOC732275), non-coding RNA	n/a	unlikely
<b>LOC400550</b>	chr16:85065632-85099967	hypothetical LOC400550 (LOC400550), transcript variant 1, non-coding RNA	n/a	unlikely
<b>FOXF1</b>	chr16:85101634-85105571	forkhead family of transcription factors, found in promoters of SPB. Paper: shows weak expression of the gene in IPF	high: lung (adult and fetal), SI	possible
<b>MTHFSD</b>	chr16:85121283-85146342	methylenetetrahydrofolate synthetase domain containing: a folate metabolizing enzyme	high: pituitary, testis,	unlikely
<b>FLJ30679</b>	chr16:85146427-85148406	hypothetical protein, function unknown	n/a	unlikely
<b>FOXC2</b>	chr16:85158358-85160038	forkhead box C2 (MFH-1, mesenchyme forkhead 1): it may play a role in the development of mesenchymal tissues.	high: liver, heart, cardiac cells	possible but unlikely
<b>FOXL1</b>	chr16:85169616-85172805	forkhead box L1 (FOXL1)	high: liver	unlikely
<b>LOC100506581</b>	chr16:85893906-85908527	hypothetical LOC100506581 (LOC100506581):	n/a	unlikely
<b>FBXO31</b>	chr16:85920443-85974895	F-box protein 31 (FBXO31): suggested to function as a tumor suppressor by generating SCF-FBXO31 complexes that target substrates critical for the normal execution of the cell cycle (i.e. ubiquitination and degradation)	high: testis, endothelial cells	unlikely
<b>MAP1LC3B</b>	chr16:85983302-85995881	microtubule-associated protein 1 light chain 3 beta: a subunit of neuronal microtubule-associated MAP1A and MAP1B proteins, which are involved in microtubule assembly and important for neurogenesis.	high: heart, brain, skeletal muscle	unlikely
<b>ZCCHC14</b>	chr16:85997353-86082961	zinc finger, CCHC domain containing 14	high: testis, SI, fetal brain	unlikely
<b>JPH3</b>	chr16:86194000-86289262	junctional protein 3: Junctional complexes b/w the plasma membrane and endoplasmic/sarcoplasmic reticulum are a common feature of all excitable cell types/ mediate cross talk b/w cell surface and intracellular ion channels.	high: whole brain, cardiac cells	unlikely
<b>KLHDC4</b>	chr16:86298919-86357099	kelch domain containing 4:	high: CD cells	unlikely



<b>SLC7A5</b>	chr16:86421130-86460601	solute carrier family 7 (cationic amino acid transporter, y+ system), member 5: involved in lung cancer	high: bronchial epithelial cells, early erythroid, pineal	unlikely
<b>CA5A</b>	chr16:86479126-86527613	carbonic anhydrase VA, mitochondrial: large family of zinc metalloenzymes that catalyze the reversible hydration of carbon dioxide. It may play an important role in ureagenesis and gluconeogenesis.	high: liver	unlikely
<b>BANP</b>	chr16:86561125-86668425	BTG3 associated nuclear protein: protein binds to matrix attachment regions. The protein forms a complex with p53 and negatively regulates p53 transcription/ functions as a tumor suppressor/cell cycle regulator	high: heart, spleen, thymus	unlikely
<b>ZNF469</b>	chr16:87021380-87034666	encodes a zinc-finger protein: Mutations in this gene cause brittle cornea syndrome- rare genetic connective tissue disorder characterized by lax joints, scoliosis and fragile sclera of the eye	ubiquitous	unlikely
<b>ZFPM1</b>	chr16:87047515-87129075	aka FOG (friend of GATA1)- GATA1 involved in erythroid differentiation	high: hematopoietic tissues	unlikely
<b>ZC3H18</b>	chr16:87164290-87225873	zinc finger CCCH-type containing 18: shown to be involved in an in vitro model of trypanosome differentiation	n/a	unlikely
<b>IL17C</b>	chr16:87232502-87234383	interleukin 17C: a T cell-derived cytokine that shares the sequence similarity with IL17. This cytokine was reported to stimulate the release of tumor necrosis factor alpha and interleukin 1 beta from a monocytic cell line.	high: activated T cells	possible
<b>CYBA</b>	chr16:87237198-87244958	cytochrome b-245, alpha polypeptide: Mutations in this gene are associated with autosomal recessive chronic granulomatous disease (CGD), that is characterized by the failure of activated phagocytes to generate superoxide, which is important for the microbicidal activity of these cells.	high: blood, lung	unlikely
<b>MVD</b>	chr16:87245849-87256996	mevalonate (diphospho) decarboxylase: catalyzes the conversion of mevalonate pyrophosphate into isopentenyl pyrophosphate in one of the early steps in cholesterol biosynthesis.	high: adipocytes, adrenal cortex	unlikely
<b>MGC23284</b>	chr16:87257282-87281095	hypothetical LOC197187 (MGC23284), transcript variant 2, non-coding RNA	n/a	unlikely
<b>SNAI3</b>	chr16:87271591-	snail homolog 3 (Drosophila): which plays roles in mesodermal	skin melanoma,	unlikely

	87280383	formation during embryogenesis	pooled lung/ spleen, lung epidermoid carcinoma,	
<b>RNF166</b>	chr16:87290404- 87300330	ring finger protein 166:	n/a	unlikely
<b>CTU2</b>	chr16:87300392- 87309287	cytosolic thiouridylase subunit 2 homolog (S. pombe): Wobble uridine tRNA thiolase Ctu1-Ctu2 is required to maintain genome integrity in S. pombe	n/a	unlikely
<b>FAM38A</b>	chr16:87309247- 87378873	family with sequence similarity 38, member A: FAM38A encodes PIEZO1, a protein that induces mechanically activated (MA) currents in various cell types; essential components of distinct mechanically activated cation channels.	high: pineal, lung, T cells	unlikely
<b>CDT1</b>	chr16:87397687- 87403167	chromatin licensing and DNA replication factor 1: involved in the formation of the pre-replication complex that is necessary for DNA replication.	high: lymphoblast cells. CD cells	unlikely
<b>APRT</b>	chr16:87403378- 87405843	adenine phosphoribosyltransferase: enzyme catalyzes formation of AMP and inorganic pyrophosphate from adenine and 5-phosphoribosyl-1-pyrophosphate; produces adenine as a by-product of the polyamine biosynthesis pathway. A homozygous deficiency in this enzyme causes 2,8-dihydroxyadenine urolithiasis.	high: liver, prostate, lung	unlikely
<b>GALNS</b>	chr16:87407643- 87450875	galactosamine (N-acetyl)-6-sulfate sulfatase: a lysosomal exohydrolase required for the degradation of the glycosaminoglycans, keratan sulfate, and chondroitin 6-sulfate. Mutations can lead to deficiencies of this enzyme, which lead to Morquio A syndrome, a lysosomal storage disorder.	high: prostate, pineal	unlikely
<b>TRAPPC2L</b>	chr16:87451007- 87455021	trafficking protein particle complex 2-like:	high: heart, liver, testis	unlikely
<b>PABPN1L</b>	chr16:87457249- 87460569	poly(A) binding protein, nuclear 1-like (cytoplasmic): important role in fertility involving the oocyte potential for embryo development.	high: oocytes	unlikely
<b>CBFA2T3</b>	chr16:87468764- 87535109	core-binding factor, runt domain, alpha subunit 2; translocated to, 3: interacts with DNA-bound transcription factors and recruit a range of corepressors to facilitate transcriptional repression. The t(16;21)(q24;q22) translocation is one of the less common karyotypic abnormalities in acute myeloid leukemia.	high: dendritic cells, CD34 cells, thymus	unlikely
<b>ACSF3</b>	chr16:87687755- 87748498	acyl-CoA synthetase family member 3 (ACSF3), nuclear gene encoding mitochondrial protein: Acyl-CoA is involved in the metabolism	n/a	unlikely

		of fatty acids.		
<b>C16orf81</b>	chr16:87753129-87757584	chromosome 16 open reading frame 81: non-coding RNA	n/a	unlikely
<b>CDH15</b>	chr16:87765664-87789401	cadherin 15, type 1, M-cadherin (myotubule): The protein is thought to be essential for the control of morphogenetic processes, specifically myogenesis, and may provide a trigger for terminal muscle cell differentiation.	high: cerebellum, skeletal muscle	unlikely
<b>ZNF778</b>	chr16:87811612-87823466	zinc finger protein 778: member of the krueppel C2H2-type zinc-finger protein family, is a candidate gene for autism and variable cognitive impairment in the 16q24.3 microdeletion syndrome.	high: liver, heart	unlikely
<b>ANKRD11</b>	chr16:87861536-88084470	ankyrin repeat domain 11: interacts with p160 nuclear receptor coactivators and inhibits ligand-dependent transcriptional activation	high: skeletal muscle, myeloid leukemia cells	unlikely
<b>SPG7</b>	chr16:88102306-88151675	spastic paraplegia 7 (pure and complicated autosomal recessive) (SPG7), nuclear gene encoding mitochondrial protein: Mutations associated with this gene cause autosomal recessive spastic paraplegia 7	high: thyroid	unlikely
<b>RPL13</b>	chr16:88154632-88157349	ribosomal protein L13: encodes a ribosomal protein that is a component of the 60S subunit. expressed at significantly higher levels in benign breast lesions than in breast carcinomas	high: epithelial cells	unlikely
<b>SNORD68</b>	chr16:88155339-88155410	small nucleolar RNA, C/D box 68	n/a	unlikely
<b>CPNE7</b>	chr16:88169677-88191155	copine VII: member of the copine family, which is composed of calcium-dependent membrane-binding proteins. contains two N-terminal C2 domains and one von Willebrand factor A domain. The protein may be involved in membrane trafficking.	high: skeletal muscle	unlikely
<b>DPEP1</b>	chr16:88207217-88232340	dipeptidase 1 (renal): a kidney membrane enzyme implicated in renal metabolism of glutathione and its conjugates	high: kidney, pancreas	unlikely
<b>CHMP1A</b>	chr16:88238345-88251630	chromatin modifying protein 1A: involved in multivesicular body sorting of proteins to the interiors of lysosomes	high: thyroid, blood, heart, lung	unlikely
<b>C16orf55</b>	chr16:88251711-88265176	chromosome 16 open reading frame 55: LOC124045	high: testis	unlikely
<b>CDK10</b>	chr16:88280577-	cyclin-dependent kinase 10: known to be essential for cell cycle	high: thyroid,	unlikely



	88290273	progression. This kinase has been shown to play a role in cellular proliferation and its function is limited to cell cycle G2-M phase.	liver, heart, PFC	
<b>SPATA2L</b>	chr16:88290266-88295622	spermatogenesis associated 2-like	n/a	unlikely
<b>C16orf7</b>	chr16:88301042-88314895	chromosome 16 open reading frame 7	high: retina, myeloid cells	unlikely
<b>ZNF276</b>	chr16:88314894-88334833	zinc finger protein 27	high: kidney, liver, lung, spleen	unlikely
<b>FANCA</b>	chr16:88331460-88410566	Fanconi anemia, complementation group A: Fanconi anemia is a genetically heterogeneous recessive disorder characterized by cytogenetic instability, hypersensitivity to DNA crosslinking agents, increased chromosomal breakage, and defective DNA repair.	high: lymphoblast cells	unlikely
<b>SPIRE2</b>	chr16:88422408-88465228	spire homolog 2 (Drosophila)	most tissues, little/none in heart and skeletal muscle	unlikely
<b>TCF25</b>	chr16:88467495-88505293	transcription factor 25 (basic helix-loop-helix): member of basic helix-loop-helix family of TFs that are important in embryonic development	high: whole brain, CD cells	unlikely
<b>MC1R</b>	chr16:88511788-88514886	melanocortin 1 receptor: encodes the receptor protein for melanocyte-stimulating hormone (MSH). Loss-of-function mutations are associated with increased pheomelanin production, which leads to lighter skin and hair color.	ubiquitous, skin	unlikely
<b>TUBB3</b>	chr16:88515918-88530006	tubulin, beta 3: primarily expressed in neurons and may be involved in neurogenesis and axon guidance and maintenance. Mutations in this gene are the cause of congenital fibrosis of the extraocular muscles type 3.	high: fetal and adult brain	unlikely
<b>DEF8</b>	chr16:88542684-88561968	differentially expressed in FDCP 8 homolog (mouse):	high: bronchial epithelial cells, smooth muscle	unlikely
<b>CENPBD1</b>	chr16:88563684-88566741	CENPB DNA-binding domains containing 1:	high: brain	unlikely
<b>AFG3L1P</b>	chr16:88566489-88590529	AFG3 ATPase family gene 3-like 1 (S. cerevisiae), pseudogene (AFG3L1P), transcript variant 2, non-coding RNA: The yeast Afg3 and	n/a	unlikely

Rca1 mitochondrial chaperones/proteases function both in posttranslational assembly and in the turnover of mistranslated or misfolded polypeptides.

<b>DBNDD1</b>	chr16:88598780-88613438	dysbindin (dystrobrein binding protein 1) domain containing 1:	high: liver, PFC	possible but unlikely
<b>GAS8</b>	chr16:88616509-88638880	growth arrest-specific 8: gene is sometimes deleted in breast and prostate cancer, is a putative tumor suppressor gene.	high: heart, skeletal muscle, lung	unlikely
<b>C16orf3</b>	chr16:88622817-88623810	chromosome 16 open reading frame 3, hypothetical. Pr. LOC750: common in breast/prostate cancer, suggests the presence of a tumor suppressor gene.	n/a	unlikely
<b>LOC100130015</b>	chr16:88636811-88641692	5-hydroxyisourate hydrolase pseudogene (LOC100130015), transcript variant 1, non-coding RNA	n/a	unlikely
<b>PRDM7</b>	chr16:88650475-88669839	PR domain containing 7: a transcription factor of the PR-domain protein family, contains a PR-domain and multiple zinc finger motifs. Transcription factors of this family: usually involved in cell differentiation, tumorigenesis.	n/a	unlikely

## Appendix C: Forward and Reverse Primer Sequences for All Genes Sequenced in Family R0942

Sequences in capital letters are located in exonic regions. Sequences in lower case letters are located in intronic and/or UTR regions.

Gene	Forward primer 		Reverse Primer 	
	primer name	sequence	primer name	sequence
EDN1	EDN1_ex1_F	attgtctggggctggaataa	EDN1_ex1_R	aacgggggaaaaatagaagca a
	EDN1_ex2_F	caggctgtgtgcttcatctg	EDN1_ex2_R	ggcacttgatgggtgttagaa
	EDN1_ex3_F	gcccagtggaatagggtgtg	EDN1_ex3_R	aacaggaaggcagtagcatga
	EDN1_ex4_F	agcctcctgaactccttct	EDN1_ex4_R	taagaaggctgtaggaaca
	EDN1_ex5a_F	caggttttgttgccaga	EDN1_ex5a_R	CAGAACTCCACCCC TGTGT
	EDN1_ex5b_F	TCCTCTGCTGGTTCCT GACT TGGCAGAAGTATTTC	EDN1_ex5b_R	AGTATGGGGGATGGA GGAAG
	EDN1_ex5c_F	CACAT	EDN1_ex5c_R	taatacagttggcccactc
SOCS1	SOCS1_ex1_F	gcggaaagagaacacaaagt	SOCS1_ex1_R	ccggagaaaggctgtgt
	SOCS1_ex2a_F	tgtgtccactgaggctgaac	SOCS1_ex2a_R	AGGCCATCTTCACGC TAAGG
	SOCS1_ex2b_F	CCCCTTCTGTAGGATG GTAGC	SOCS1_ex2b_R	AGGGGAAGGAGCTCA GGTAG
	SOCS1_ex2c_F	AGAGCTTCGACTGCCT CTTC	SOCS1_ex2c_R	AGGTAGGAGGTGCGA GTTCA
	SOCS1_ex2d_F	GCAGACCCCTTCTCAC CTCT	SOCS1_ex2d_R	cagcaggctagccttaggac
TELO2	TELO2_ex1_F	atggggagaaactaaggcgaga	TELO2_ex1_R	caacagcttacggggaacac
	TELO2_ex2_F	agcgaactctgggtggaac	TELO2_ex2_R	aggctctgggtcagtcctt
	TELO2_ex3_F	tggacttcacgctgttatg	TELO2_ex3_R	gacgggatcaaccagagact
	TELO2_ex4-5_F	agttttgctgcgttgag	TELO2_ex4-5_R	acacctgtggcacctacat
	TELO2_ex6_F	agaggctcgtgcttagaat	TELO2_ex6_R	ctcactcagagccgtcacag
	TELO2_ex7-8-9_F	agtctgcgcatggctctg	TELO2_ex7-8-9_R	gagctcagggtgagaacct
	TELO2_ex10-11_F	tggcctgaggtggaatctta	TELO2_ex10-11_R	agagaagaagcagggagcag
	TELO2_ex12-13_F	ctgctcgtgcttctct	TELO2_ex12-13_R	agccatatgtcacctcagtg
	TELO2_ex14-15_F	gtagctccctcaatgccatc	TELO2_ex14-15_R	acaaggagaccagggttttg
	TELO2_ex16_F	ctgtgagctacgggaagt	TELO2_ex16_R	ctgggtctcctacagccaaa
	TELO2_ex17_F	gctgctttgggactctct	TELO2_ex17_R	gttcacaccaaaccactc

	TELO2_ex18-19_F TELO2_ex20_F TELO2_ex21a_F TELO2_ex21b_F	ccggaaatgtctgctcctac aggcagacacaggggtcttat acagtgctgcctgcacagag GCTGCTTCTGCAGAGACTCAA	TELO2_ex18-19_R TELO2_ex20_R TELO2_ex21a_R TELO2_ex21b_R	ctgcctggcacacttctctat aggagctgcaacacacacac TAGCGGACTGGCTGC ACTAT gctccccacacatcatatcc
CD83	CD83_ex1_2_F CD83_ex3_F CD83_ex4_F CD83_ex5a_F CD83_ex5b_F CD83_ex5c_F CD83_ex5d_F	caccctctcggaatctggtt ttccgaagatgtagccttg ctgatggtgggaagaggaga gagacgccagtgaatggtt TGAAGATGGCATCCTGTGAA GCTGTACCAGCCCAGATGTT TGTTGCATGGGCTAATGAAG	CD83_ex1_2_R CD83_ex3_R CD83_ex4_R CD83_ex5a_R CD83_ex5b_R CD83_ex5c_R CD83_ex5d_R	ggagatggtctcgtgtct tccacttcagccacagatg cctcaaaaagtccaggggttc AATGTCCAGGAGGTTGACCA AGACCCTGCTGGGGA ATAGTTGGATAGCACCCTCTCATCC gcagaagcatctctgggaag
DTNBP1	DTNBP1_ex1_F DTNBP1_ex2_F DTNBP1_ex3_F DTNBP1_ex4_F DTNBP1_ex5_F DTNBP1_ex6_F DTNBP1_ex7_F DTNBP1_ex8_F DTNBP1_ex9_F	aggagctactggccctctc gctggagcagacagagtga aaagcatgtgaccagattca agcaccacaggatattccac attggaggttctgcactg ttatggtttaatgttgccagt ccactgtgggcatttaaagta gcgagcagctgtttataccc aatgccaatggaagttcacc	DTNBP1_ex1_R DTNBP1_ex2_R DTNBP1_ex3_R DTNBP1_ex4_R DTNBP1_ex5_R DTNBP1_ex6_R DTNBP1_ex7_R DTNBP1_ex8_R DTNBP1_ex9_R	tctcacatgactccctcca gcactcaacaaatctaaggttca aattccaatcactgcattaaga gggagtttcataatcaacgact ccagaattcatgtgttctga cagccgggacttctaagag ctgcacctcctaacacacc acagaacgggaactctgagga agatggttctcacgtctcacc

	DTNBP1_ex1 0_F	gtggcggcaattctgatt	DTNBP1_ex10 _R	tgacgctccctcactctaa
VEGFA	VEGFA_ex1a_F	ggcaaaagtgagtacactgct CCAACTTCTGGGCTGT	VEGFA_ex1a_R	GGCCCGAGCTAGCAC TTCT
	VEGFA_ex1b_F	TCTC	VEGFA_ex1b_R	ctgcacctaagacgacaga
	VEGFA_ex2_F	aagacttgagagaagccagagg	VEGFA_ex2_R	aattaggccatccaccatc
	VEGFA_ex3_F	tggaaggactgcctgattc	VEGFA_ex3_R	ccacctgttccaaagtgtt
	VEGFA_ex4_F	ggtgtgccatctgggtatg	VEGFA_ex4_R	atgcttaaccctggcacaga
	VEGFA_ex5_F	tcatcaccatcttaacccttc	VEGFA_ex5_R	cccaacagaggtagccaaga
	VEGFA_ex6_F	CTGTGTGGCTTTGCTTTGGT C	VEGFA_ex6_R	TCTACCCGTTGGTGCCAATT A
	VEGFA_ex7_F	CTAGCCAGTGTGCCTCTTT C	VEGFA_ex7_R	TGTGATGCCCTCTCTGACT T
	VEGFA_ex8_F	ctctcacttggccctaacc	VEGFA_ex8_R	tctgtcatggtgatggtgt
SNRNP25	SNRNP25_ex1_F	cacttgcattggtcactgc	SNRNP25_ex1_R	agtctcccaccagcacctt
	SNRNP25_ex2_F	cggctgttcttgtgttg	SNRNP25_ex2_R	caggctgaccagcactgtta
	SNRNP25_ex3_F	gcacagagggtgtggttct	SNRNP25_ex3_R	agagctgtttccagggtggtg
	SNRNP25_ex4_F	cagggtccataggactgcat	SNRNP25_ex4_R	cttggaggagggtgttatct
	SNRNP25_ex5a_F	tgtgtttctctggtctctc	SNRNP25_ex5a_R	CCCAGGTTACTGTGG CTTTC
	SNRNP25_ex5b_F	TTGGCCTGAGACTGAC CTCT	SNRNP25_ex5b_R	agggaaaccttaagggtgag
MSLN	MSLN_ex1_F	gacaggaggagccagtcag	MSLN_ex1_R	acaggcactccaggagagaag
	MSLN_ex2_F	gtacatgggcctgagccact	MSLN_ex2_R	accgtggatacgtgacagaga
	MSLN_ex3_F	gggaactcctgctccagaga	MSLN_ex3_R	aggtggagcagagagggaag
	MSLN_ex4_F	atctgactgggctcaggact	MSLN_ex4_R	gagctgcctgtgaattctc ATCTGGGctgtgaagagag
	MSLN_ex5_F	tccacttccagattctcg	MSLN_ex5_R	g
	MSLN_ex6_F	caggccacagcagagatgt CTGCACCCGTTTCTTC	MSLN_ex6_R	ctcagcgcttttggaaactg
	MSLN_ex7_F	TCC	MSLN_ex7_R	gtgtcctcacgctgctgata
	MSLN_ex8-9_F	ctggcctctccctctctg	MSLN_ex8-9_R	cactggcaaacagggtattct



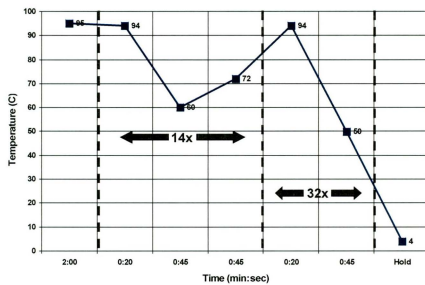
	MSLN_ex10_ F	tcattctcacaggagggtg	MSLN_ex10_ R	GTAGAGctcaggccaaga g
	MSLN_ex11_ F	ctgctgtgtccaagccatc	MSLN_ex11_ R	ctccaccaccagcataacct
	MSLN_ex12- 13_F	gtcatgtggcatgagattgg	MSLN_ex12- 13_R	ccgagtgaggtctgtgcttc
	MSLN_ex14_ F	ctgtaaggcaagtgggcttc	MSLN_ex14_ R	caccagggtcttccactt
	MSLN_ex15_ F	ggcagtttggacacaggaga	MSLN_ex15_ R	CAGGACGGTGAGAAC AGGTC
	MSLN_ex16_ F	CCTGGTCCTAGACCTC AGCA	MSLN_ex16_ R	agcagggtcaggaagacctc
IL17C	IL17C_Ex1_F	ttcccaggagggaagtgg	IL17C_Ex1_R	cctccctacccccagact
	IL17C_Ex2_F	agcctctctgggcttcagtt	IL17C_Ex2_R	tgacaggtgagaaccaccaa
	IL17C_Ex3a_ F	ctgcctcaggtctctcctg	IL17C_Ex3a_ R	TGTTGGGGGAGGCAT ATAAA
	IL17C_Ex3b_ F	GGGCCCCCTAGACTGG ACAC	IL17C_Ex3b_ R	gctgcgtagacctttctgga

Primer sequences for validating *MUC5B* rs35705950 genotypes:

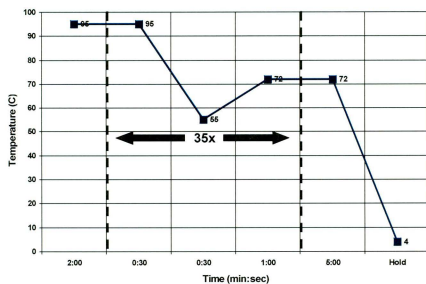
MUC5B_F	ggccagaatgaggacagt
MUC5B_R	gtttgctcagcgtgtttgaa

## Appendix D: Thermocycler programs used

### "Touchdown" Thermocycler Program

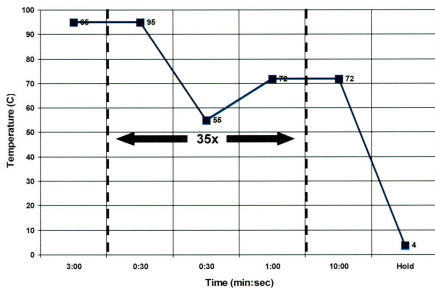


### "Standard" Thermocycler Program

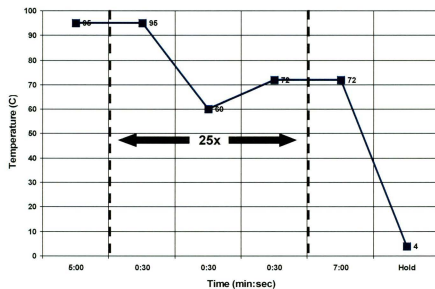


For "Standard" programs, the 55°C temperature was changed for some conditions and named accordingly. For example, if 55°C was changed to 57°C, the condition was named "Standard57".

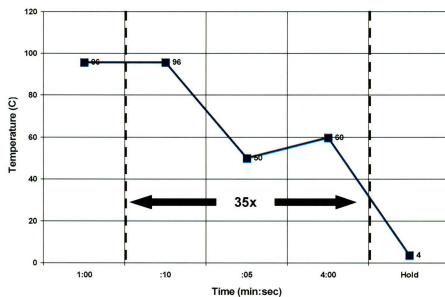
### "AP Standard" Thermocycler Program



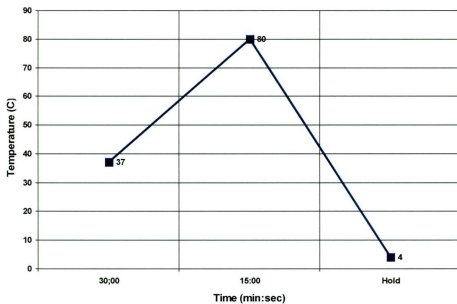
### "Gal12new" Thermocycler Program



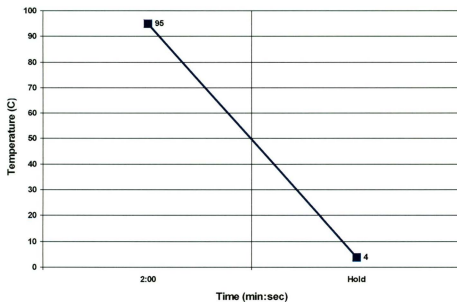
### "Abiseq" Thermocycler Program



### "Exosap" Thermocycler Program



### "Denature" Thermocycler Program



## Appendix E: Setting up plate information for genotyping using SDS version 2.4 software

- Once the SDS version 2.4 software is opened, input the DNA numbers for the plate into both curves: a Standard Curve (AQ- Allelic Quantification) and an Allelic Discrimination Curve (AD). Save both files.
- Open up SDS 2.4 and open up the AD file that was created. Click on the "Instrument" tab and click "Connect to Instrument" so the machine and computer can connect. Once connected, click the "Open/Close" button. The door will slowly open and a tray will come out. Place the prepared 0.1ml MicroAmp Fast Optical 96 reaction well PCR plate on the tray with the correct orientation. Click the "Open/Close" button again and wait for the plate to go into the machine.
- **PRE-READ in AD:** Within the AD file that is open, highlight all of the samples and click the "Pre-Read" button. Wait for the Pre-Read to take place. Save and close the AD file.
- **AQ RUN:** Open up the AQ file and highlight all samples in the file. Set up the thermocycler conditions to the desired condition. Choose "Fast" or "Standard" and choose the correct thermocycler condition under the "Thermal Cycler" tab (conditions vary based on the PCR master mix used). Save the file and click the "Start Run" button. This run will take approximately 90 minutes. Once the AQ Run is completed, save the AQ file again.
- **POST-READ in AD:** Open the AD file and do a "Post-Read." This will allow for analysis of the results. Save the file after the Post-Read.
- **ANALYSIS:** In the AD file, press the green "play" button in the icons at the top of the program to begin the analysis. Save the analysis and transfer the files to a computer. Analyze by saving a copy of the AD plot as an image file and exporting the data as a .txt file. Open the .txt file in Microsoft Excel to analyze it in table form.

## **Appendix F: Steps for Ethanol precipitation of cycle sequencing reactions**

1. Add 5 $\mu$ l of 125mM EDTA to each reaction well.
2. Add 65 $\mu$ l of 100% EtOH to each well, cover, vortex.
3. Allow to precipitate 2 hours to overnight in the dark at room temperature.
4. Centrifuge plate for 30 minutes at 3000 x g.
5. Remove plate and invert plate gently to decant EtOH.
6. Keep plate inverted and place on folded paper towel on plate carrier and spin up to 200rpm and stop.
7. Add 150 $\mu$ l of 70% EtOH wash.
8. Centrifuge plate for 15 minutes at 3000 x g.
9. Remove plate and invert plate gently to decant EtOH.
10. Keep plate inverted and place on folded paper towel on plate carrier and spin up to 200rpm and stop (same as step 6).
11. Allow the plate to dry completely (uncapped) for ~30 minutes in the dark.
12. Add 15-18 $\mu$ l of deionised formamide (Hi-Di formamide) to each reaction well.
13. Vortex to resuspend and centrifuge up to 600rpm for a few seconds.
14. Run 'denature' program on thermocycler and quick-chill before loading on sequencer.

**Appendix G: All genes fully sequenced in the regions of interest on chromosome 6 and 16.**

	Previously sequenced* in at least 2 affected patients from 6 FPF families		Sequenced in 5 affected and 2 unaffected individuals in Family R0942	
Chromosome	6	16	6	16
Genes sequenced	<i>TFAP2a</i> <i>TFAP2β</i> <i>WRNIP1</i> <i>SERPINB1</i>	<i>DEXI</i> <i>EMP2</i> <i>VASN</i> <i>GLIS2</i> <i>TID1</i> <i>PPL</i>	<i>EDN1</i> <i>CD83</i> <i>DTNBP1</i> <i>VEGFA</i>	<i>SOCS1</i> <i>TELO2</i> <i>SNRNP25</i> <i>MSLN</i> <i>IL17C</i>

\*For more detail on previously sequenced candidate genes, review Kamel, 2010.





# THORAX

An International Journal of Respiratory Medicine

## Association between a Promoter SNP in MUC5B and Idiopathic Pulmonary Fibrosis in the Newfoundland population

Journal:	Thorax
Manuscript ID:	Draft
Article Type:	Research Letter
Date Submitted by the Author:	n/a
Complete List of Authors:	Pirzada, Ashar; Memorial University of Newfoundland, Discipline of Genetics Mahoney, Krista; Memorial University of Newfoundland, Discipline of Genetics Zhai, Guangju; Memorial University of Newfoundland, Discipline of Genetics Noble, Barbara; Eastern Health, Fernandez, Bridget; Memorial University of Newfoundland, Discipline of Genetics Woods, Michael; Memorial University of Newfoundland, Discipline of Genetics
Keywords:	Idiopathic pulmonary fibrosis

SCHOLARONE<sup>®</sup>  
Manuscripts

## **Association between a Promoter SNP in *MUC5B* and Idiopathic Pulmonary Fibrosis in the Newfoundland population**

Ashar Pirzada<sup>1</sup>, Krista Mahoney<sup>1</sup>, Guangju Zhai<sup>1</sup>, Bridget A. Fernandez<sup>1,2,3</sup>, Michael O. Woods<sup>1</sup>.

*Discipline of <sup>1</sup>Genetics, <sup>2</sup>Medicine, Memorial University of Newfoundland, St John's, NL, Canada; <sup>3</sup>Provincial Medical Genetic Program, Eastern Health, St. John's, NL, Canada*

Ashar Pirzada: [ashar.pirzada@med.mun.ca](mailto:ashar.pirzada@med.mun.ca); Krista Mahoney: [kmahoney@mun.ca](mailto:kmahoney@mun.ca);  
Guangju Zhai: [guangju.zhai@med.mun.ca](mailto:guangju.zhai@med.mun.ca); Bridget A. Fernandez:  
[bfernandez@nl.rogers.com](mailto:bfernandez@nl.rogers.com); Michael O. Woods: [mwoods@mun.ca](mailto:mwoods@mun.ca)

### **Corresponding author:**

Michael O. Woods  
Discipline of Genetics  
Health Sciences Centre, Rm 4333  
300 Prince Philip Drive  
St. John's, Newfoundland, Canada A1B 3V6  
Tel: 709-777-7334; Fax: 709-777-7497  
Email: [mwoods@mun.ca](mailto:mwoods@mun.ca)

**Keywords:** pulmonary fibrosis, MUC5B, Newfoundland

## ABSTRACT

**Background:** Recently, a promoter variant (rs35705950) upstream of *MUC5B* has been shown to be associated with IPF in US populations.

**Methods:** The SNP rs35705950 was genotyped by a TaqMan SNP Genotyping assay. A case-control analysis was carried out using 110 affected individuals and 277 healthy controls from the Newfoundland population.

**Results:** There is significant association between rs35705950 genotypes and IPF. The odds ratios for all individuals affected with IPF who were heterozygous and homozygous for the variant allele of this promoter polymorphism, respectively were 5.4 (95% CI 3.3 to 9.6,  $P < .001$ ) and 12.2 (95% CI 3.3 to 44.7,  $P < .001$ ). Furthermore, a large family displayed segregation of the variant allele with the phenotype.

**Conclusions:** The minor T allele of rs35705950 is significantly associated with PF in this Newfoundland cohort and may contribute to a small proportion of FPF families with unknown genetic etiology.

Recently, Seibold et al.<sup>1</sup> identified a variant (rs35705950) located in a promoter region 3 kb upstream of a major gel-forming mucin gene, *MUC5B*, that is associated with both sporadic and familial IPF in US populations. Our study validates the association of rs35705950 with both sporadic and familial forms of IPF in the Newfoundland population and we also show segregation of the variant in a large FPF pedigree.

One-hundred and ten affected individuals were recruited through respiratory clinics in the only tertiary care hospital in Newfoundland, Canada. All probands were sequenced to exclude possible IPF-causing genetic variations in *TERT*, *TERC*, *SFTPC* and *SFTPA2*.<sup>2</sup> Sixty-eight of the PF patients were sporadic and 42 were familial, with at least one affected first or second-degree relative. All affected individuals had high resolution computed tomograms of the chest and met the ATS/ERS criteria for diagnosis of IPF.<sup>3</sup> Healthy individuals, previously recruited for a colorectal cancer study by random-digit-dialing, served as a control cohort<sup>4</sup>. To genotype rs35705950, a TaqMan SNP Genotyping assay using a 7900HT Real-time PCR analyzer (Applied Biosystems, CA, USA) was implemented. A case-control analysis was carried out using 110 affected individuals and 277 healthy controls. The analysis was performed twice to determine odds ratios and 95% confidence intervals in the two groups - once with all 110 cases, while controlling for familial samples via clustering; and once by using only the 80 unrelated familial and sporadic probands (SPSS v20.0.0, IBM Software).

The clinical variables and results for both analyses are summarized in Table 1. There was a significant association between rs35705950 genotypes and IPF. The odds ratio for individuals affected with IPF who were heterozygous for the minor allele was 5.4 (95% confidence interval, 3.3 to 9.6,  $P < .001$ ). The odds ratio for individuals affected with IPF who were homozygous for the minor allele was 12.2 (95% confidence interval, 3.3 to 44.7,  $P < .001$ ). Interestingly, the unrelated proband analysis showed the odds ratio for

affected probands homozygous for the variant T allele was slightly higher than in the same analysis for all cases (OR = 15.2; 95% CI, 4.3 to 52.6,  $P < .001$ ).

Our cohort included affected probands and relatives from 12 PF families. Segregation analysis was performed on the 12 families. Family R0942 exhibited segregation of the variant T allele (Figure 1). In this family, all seven affected individuals were heterozygous for the variant T allele. To determine if segregation of the risk allele occurred by chance, SISA (simplified rapid segregation analysis) was conducted<sup>5</sup>. From this analysis, the probability that co-segregation of the variant T allele with PF occurred by chance was 1.56% in this family.

Our results further support a significant association between the minor T allele of rs35705950 and PF. Interestingly, the minor T allele segregated, with reduced penetrance, in an autosomal dominant manner in PF families. These findings further support that the minor T allele of rs35705950 is associated with developing PF in sporadic patients. It is also the first study to show that this allele may act as a relatively highly penetrant allele in some families.

**Table 1: Clinical variables and findings for all cases, probands and controls.**

Clinical variables	All cases (n=110)	Cases: probands only (n=80)	Controls (n=277)
Age (year)*	62.1 ± 12.3	63.1 ± 9.6	61.2 ± 9.4
Gender	M = 67 (60.9%) F = 43 (39.1%)	M = 51 (63.8%) F = 29 (36.2%)	M = 158 (56.5%) F = 112 (41.5%)
Smoking status**	smoker = 17 (15.5%) ex-smoker = 73 (66.3%) occasional = 3 (2.7%) never = 17 (15.5%)	smoker = 6 (7.5%) ex-smoker = 62 (77.5%) occasional = 0 (0.0%) never = 12 (15.0%)	smoker = 33 (12.2%) ex-smoker = 128 (47.4%) occasional = 2 (0.01%) never = 107 (39.6%)
Smoking duration (years)	25.0 ± 16.5	25.0 ± 16.0	26.1 ± 15.6
Pulmonary function tests: mean % of predicted value***			
FVC	84.1% ± 18.8%	83.2% ± 18.1%	N/A
FEV1	83.0% ± 15.9%	82.1% ± 15.8%	N/A

TLC	81.1% $\pm$ 16.6%	81.4% $\pm$ 15.8%	N/A
DLCO	59.4% $\pm$ 17.5%	56.6% $\pm$ 14.8%	N/A
Diagnosis by lung biopsy	45 (40.9%)	35 (43.8%)	N/A
Allele frequency:			
Wild type allele (G) frequency	63.60%	62.50%	87.40%
Minor allele (T) frequency	36.40%	37.50%	12.60%
Genotype:			
G/G	39 (35.4%)	28 (35.0%)	211 (76.2%)
G/T	62 (56.4%)	44 (55.0%)	62 (22.4%)
T/T	9 (8.2%)	8 (10.0%)	4 (1.4%)
Odds Ratio (95% CI)			
G/T vs. G/G	5.4 (3.0 - 9.6), $P < .001$	5.3 (3.1 - 9.2), $P < .001$	N/A
T/T vs. G/G	12.2 (3.3 - 44.7), $P < .001$	15.2 (4.3 - 52.6), $P < .001$	N/A
Chi-squared, P value for Hardy-Weinberg equilibrium	$\chi^2 = 5.22$ , $P = .02$	$\chi^2 = 2.40$ , $P = .12$	$\chi^2 = .05$ , $P = .82$

Control population age, gender, smoking status and duration is based on n=270 since questionnaires were not returned by 7 participants. Control population allele frequency, genotypes and odds ratio based on n=277.

\*For cases, age of diagnosis is used. For controls, age of sample collection is used.

\*\*For smoking status: Ex-smoker (all cases): smoking duration is  $28.9 \pm 13.8$  years with 71.2% having smoked over 20 years. Ex-smoker (probands only): smoking duration is  $28.7 \pm 13.1$  years with 71.0% having smoked over 20 years. Ex-smoker (controls): smoking duration is  $22.8 \pm 14.8$  years with 53.1% having smoked over 20 years. 5 total occasional smokers: only smoke socially with 1/5 having smoked over 20 years.

\*\*\*abnormal if FVC, FEV1, TLC or DLCO = <80% predicted

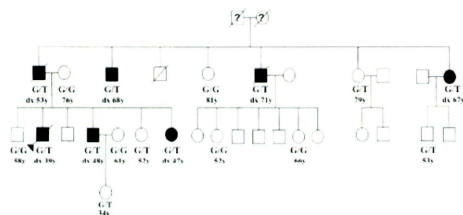
FVC = forced vital capacity, FEV1 = forced expiratory volume exhaled in the 1st second, TLC = total lung capacity, DLCO = diffusing capacity of the lung for carbon monoxide.

## ACKNOWLEDGEMENTS

We would like to express our gratitude to the patients and families who participated in this project. We also thank Dr. Sevtap Savas and Amit Negandhi for their technical assistance and Dr. Roger Green for his assistance with control recruitment.

## REFERENCES

1. Seibold MA, Wise AL, Speer MC, *et al.* A common MUC5B promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011;364:1503-12.
2. Fernandez BA, Fox G, Bhatia R, *et al.* A Newfoundland cohort of familial and sporadic idiopathic pulmonary fibrosis patients: clinical and genetic features. *Respir Res* 2012;13:64.
3. American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS board of directors, June 2001 and by the ERS Executive Committee, June 2001. *Am J Respir Crit Care Med* 2002;165:277-304.
4. Woods MO, Younghusband HB, Parfrey PS, *et al.* The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut* 2010;59:1369-77.
5. Moller P, Clark N, Maehle L. A Simplified method for Segregation Analysis (SISA) to determine penetrance and expression of a genetic variant in a family. *Hum Mutat* 2011;32:568-71.



**Figure 1.** Family R0942 with *MUC5B* variant genotypes (chr. 11: g.1241221G>T). The minor "T" allele is found in all affected individuals. Arrowhead indicates proband. Dx indicates age at diagnosis. Current age is provided below unaffected individuals. No clinical records available for the founding couple.

Figure 1. Family R0942 with *MUC5B* variant genotypes (chr. 11: q.1241221G>T). The minor "T" allele is found in all affected individuals. Arrowhead indicates proband. Dx indicates age at diagnosis. Current age is provided below unaffected individuals. No clinical records available for the founding couple.

352x264mm (72 x 72 DPI)



## Appendix J: Publisher's permissions to use copyright materials

This is a License Agreement between Ashar Pirzada ("You") and Elsevier ("Elsevier"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

License Number	3254820435021
License date	Oct 23, 2013
Licensed content publisher	Elsevier
Licensed content publication	The American Journal of Human Genetics
Licensed content title	Genetic Defects in Surfactant Protein A2 Are Associated with Pulmonary Fibrosis and Lung Cancer
Licensed content author	Yongyu Wang, Phillip J. Kuan, Chao Xing, Jennifer T. Cronkhite, Fernando Torres, Randall L. Rosenblatt, J. Michael DiMaio, Lisa N. Kinch, Nick V. Grishin, Christine Kim Garcia
Licensed content date	9 January 2009
Licensed content volume number	84
Licensed content issue number	1
Number of pages	8
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	INHERITED PREDISPOSITION TO IDIOPATHIC PULMONARY FIBROSIS IN THE NEWFOUNDLAND POPULATION
Expected completion date	Dec 2013
Estimated size (number of pages)	145
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD

This is a License Agreement between Ashar Pirzada ("You") and Nature Publishing Group ("Nature Publishing Group"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

License Number	3252040503197
License date	Oct 18, 2013
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Cancer
Licensed content title	Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins
Licensed content author	Paula Martínez, María A. Blasco
Licensed content date	Feb 24, 2011
Type of Use	reuse in a thesis/dissertation
Volume number	11
Issue number	3
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	The various functions of telomerase
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	INHERITED PREDISPOSITION TO IDIOPATHIC PULMONARY FIBROSIS IN THE NEWFOUNDLAND POPULATION
Expected completion date	Dec 2013
Estimated size (number of pages)	145
Total	0.00 USD

To Whom It May Concern:

This letter shall confirm that I, Fady Kamel, am the author of the thesis titled Determining the Genetic Etiology of Familial Pulmonary Fibrosis in Six Newfoundland Families.

I have granted Ashar Pirzada permission to reuse tables, figures and material from my thesis to supplement his Masters thesis, Inherited Predisposition of Idiopathic Pulmonary Fibrosis in the Newfoundland Population.

Sincerely,



Fady Kamel  
B.Sc. (hon), M.Sc.